(continued from part 35)

Reproduction methods

There are many techniques available for the conversion of an electrical signal into permanent marks on paper or film. They all share a scanning mechanism similar to that used at the transmittingterminal, together with circuits for synchronising the scan at the receiver with the scan at the transmitter so that the reproduced points correspond exactly to those of the original image.

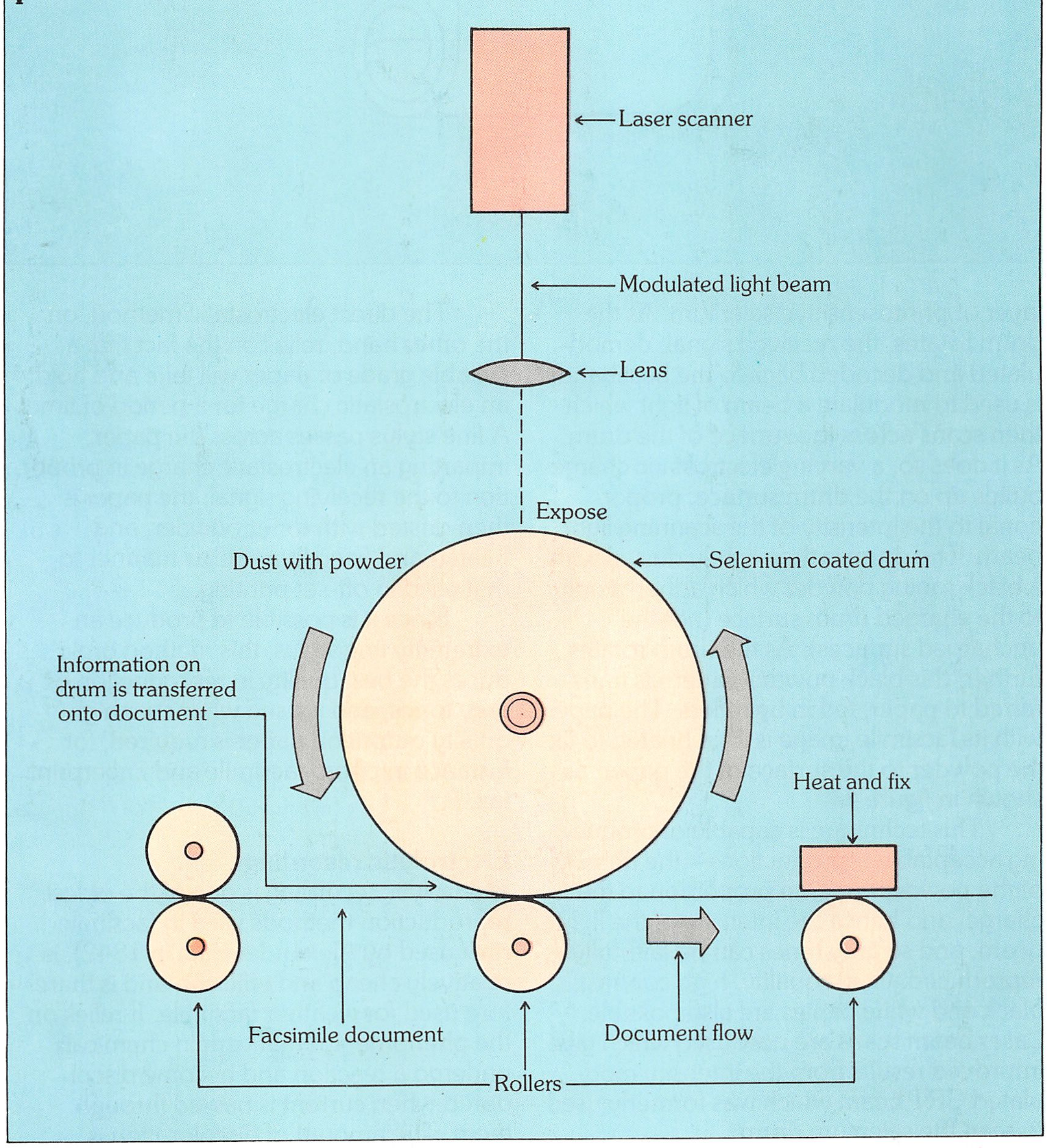
The various reproduction techniques differ in their advantages and disadvantages in terms of cost and convenience and also in their ability to generate grey tones and fineness of detail. As with scanning and sensing methods, each reproduction method is suited to a particular application.

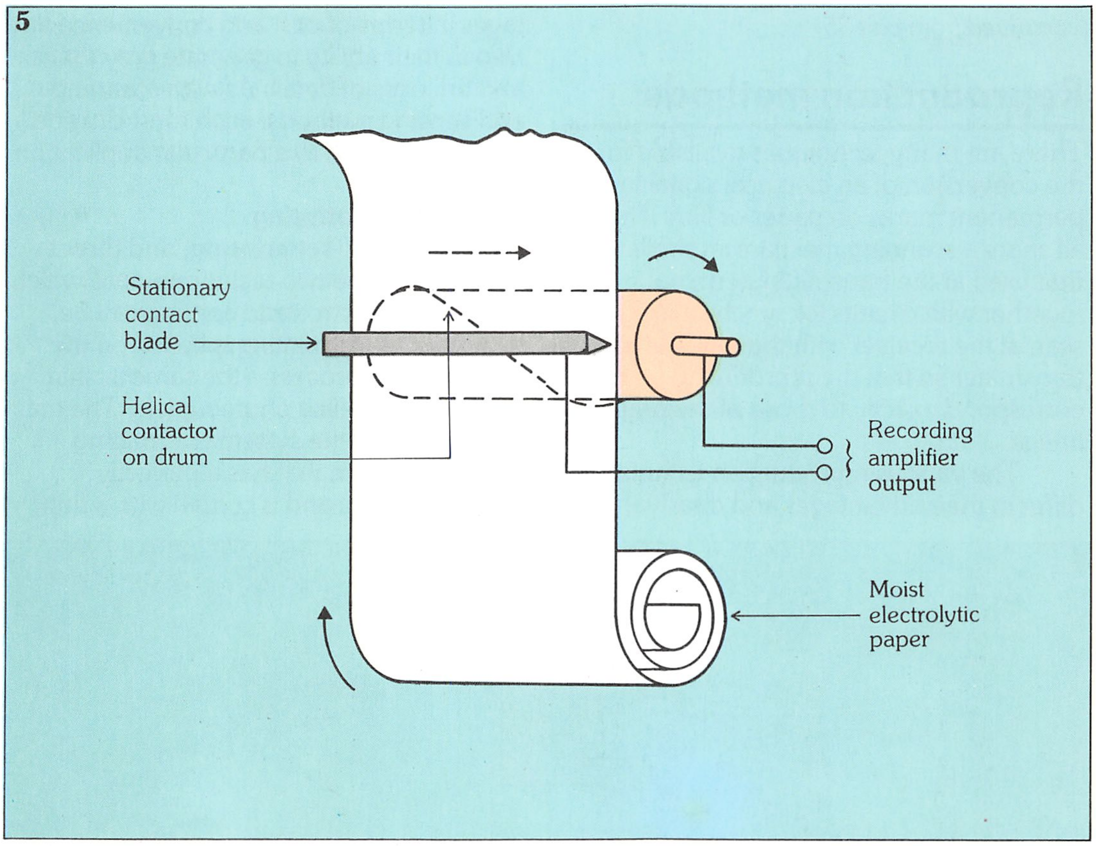
Electrostatic printing

Transfer, or **off-set printing,** and **direct printing** are the two techniques into which the area of electrostatic printing can be divided. Off-set printing is based on the **xerographic** process – the same as that used in most office photocopiers. The main component of the system is a rotating cylindrical drum that has a perfectly smooth surface and is coated with a thin

4. The off-set printing method is based on the xerographic process commonly used in office photocopiers.

4





5. Electrolytic printing involves the use of a special paper saturated with electrolyte. The diagram shows the helix and blade technique on which electrolytic printers are now based.

layer of photosensitive selenium. As the drum rotates, the received signal, demodulated and decoded back to the baseband, is used to modulate a beam of light which then scans across the surface of the drum. As it does so, a varying electrostatic charge builds up on the drum surface, proportional to the intensity of the scanning light beam. The drum is then lightly dusted with a black toning powder which adheres only to the charged drum surface (not the uncharged surfaces). As the drum rotates further, this black powder pattern is transferred to paper, fed in by rollers. The paper with its facsimile image is then heated to fix the powder to the surface of the paper, as shown in *figure 4*.

This technique is capable of producing acceptable reproductions – the density of the powder varies in proportion to the charge, and hence the intensity of the light beam, and so grey tones can be faithfully reproduced. Good quality, high contrast, black and white copies are also possible. Laser beam scans are now used which give improved results from the intensity modulated CRT beam which was formerly used to scan the selenium drum.

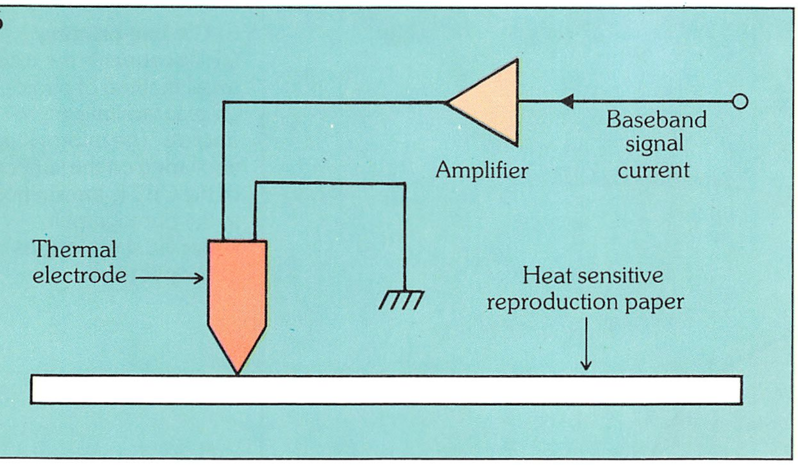
The direct electrostatic method, on the other hand, relies on the fact that a suitable grade of paper will take and hold an electrostatic charge for a period of time. A fine stylus passes across the paper, imparting an electrostatic charge in proportion to the receiving signal; the paper is then dusted with toner powder, and heated and fixed in a similar manner to that used in off-set printing.

Since it is possible to produce an extremely fine stylus, this method produces the best quality in reproduction of grey tones, and is used whenever high quality output on paper is required, for instance in photo facsimile and fingerprint fax.

Electrolytic recording

Electrolytic recording is one of the oldest reproduction methods used in facsimile (first used by Alexander Bain in 1842), is relatively cheap and efficient, and is therefore used for weather facsimile. It relies on the phenomenon that certain chemicals undergo a reaction and become discoloured when current is passed through them. The amount of discoloration is

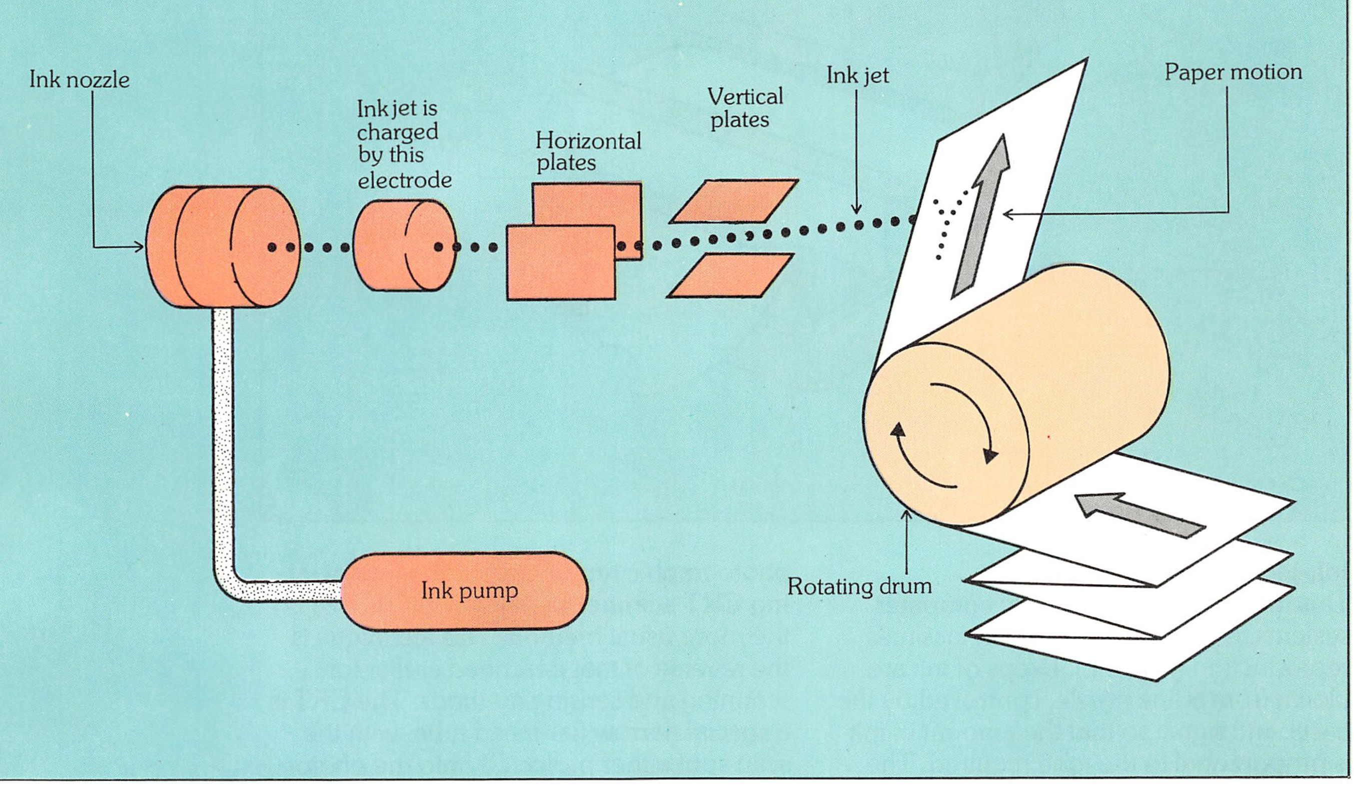
6



not corrode or tarnish in air or water nor be attacked easily by acids). The helix and the blade make contact at just one point, but as the drum rotates, the point sweeps along the axis of the drum, effectively scanning one line per revolution.

The stationary blade is connected to the amplified baseband signal, and as the drum rotates a current proportional to the signal flows from the blade to the helix, passing through the wet paper sandwiched between them. Thus, as the drum completes one revolution, a line of varying darkness appears. The paper is then dried

7



6. Electrothermal techniques rely on a paper which is dry coated with a current sensitive powder.

7. The ink-jet printing process is occasionally used in facsimile reproduction. The ink is controlled by a baseband signal to give an amount proportional to the tone required.

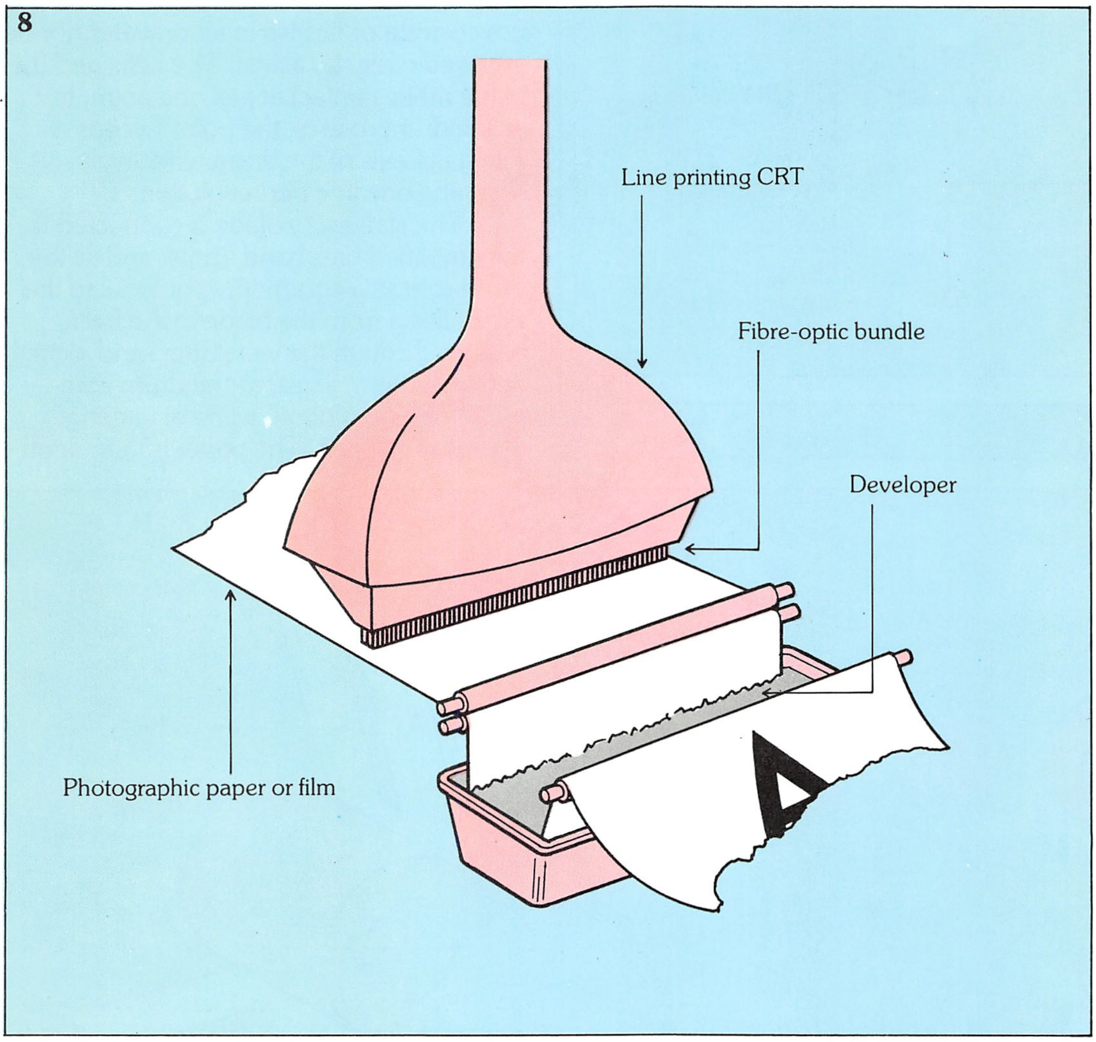
roughly proportional to the current, but cannot be relied upon to produce high quality reproductions.

Flat-bed scanning systems were once used in electrolytic printers but most of these have now been replaced by the helix and blade technique shown in *figure 5*. A special paper, saturated with the electrolyte (electrolyte printers are sometimes called **wet printers** because of this) passes between the surface of a rotating drum and a stationary blade. On the drum, there is one turn of a helix formed from a noble metal (e.g. gold, silver, platinum – one that will

and separated from the roll with a cutting blade.

Electrothermal printing

This also relies on a specially coated paper, but one which is dry coated with a current sensitive powder which decomposes to produce a mark proportional to the current flow. A moving stylus on a flat-bed scanner is normally used in electrothermal printing (*figure 6*). It is capable of producing high contrast black and white copies, but is not suited to continuous tone reproduction, such as a photograph.



8. The line-printing CRT scanner is the most usual method of photographic facsimile reproduction. The information appearing on the surface of the CRT is transmitted to the photographic paper via short lengths of fibre optic cable.

Ink-jet printers

This method, borrowed from computer systems, is occasionally used in facsimile reproduction (*figure* 7). Drops of ink are blown from a fine nozzle, controlled by the baseband signal so that the amount of ink is proportional to the tone required. The scanning process can also be controlled electronically, because the fine drops of ink are given an electric charge and then scanned by electrostatic deflection plates in the same fashion that an electron beam is scanned in a CRT.

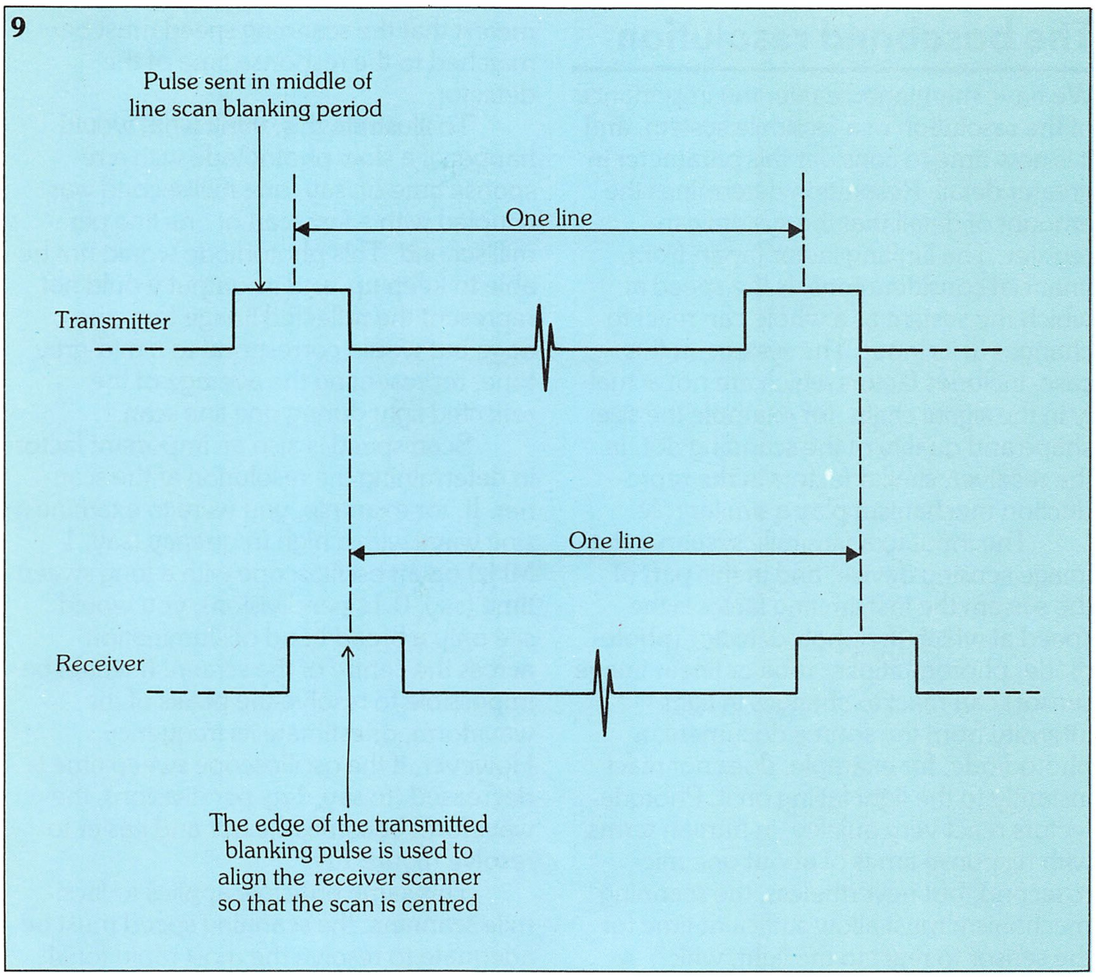
Photographic reproduction

Facsimile signals are frequently reproduced onto photographic film – either positive or negative, as required. Newspaper pages sent by facsimile, for example, are reproduced onto film that can be used directly as the printing master. Although laser scanning techniques can be used for photographic reproduction, the **line-printing CRT scanner** (as shown in *figure 8*) is the more usual method. This technique is the reverse of that described earlier for scanning and sensing methods. The CRT is a special narrow flat-faced tube, with the scan spot either projected onto the photographic negative, or fed via short lengths of fibre optic cable (methods similar to this are also used in phototypesetting).

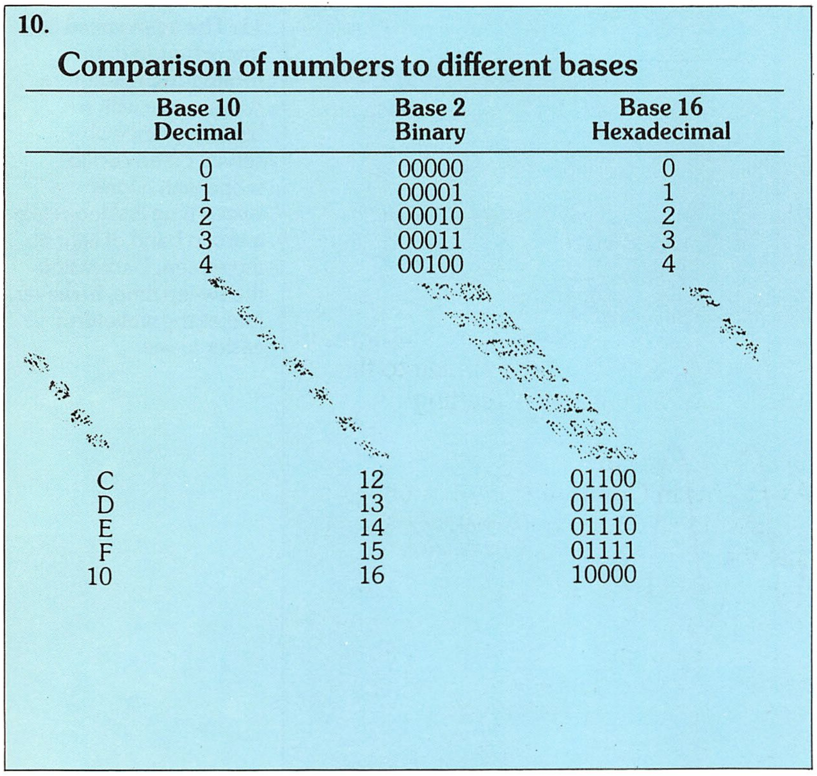
Synchronisation

As we saw in *Communications 4* when discussing television broadcasting, it is vital that the transmitting and receiving equipment should stay in step. During facsimile transmission, this means that the printer must start scanning the first line at the same point as the scan starts at the transmitting end, and the two should remain synchronised all the way down the page.

With modern electronic techniques it



9. In facsimile transmission, it is vital that the transmitter and the receiving printer should be accurately synchronised. This diagram shows how an initial signal is used to align the receiver scanner.



10.

Comparison of numbers to different bases

Base 10 Decimal	Base 2 Binary	Base 16 Hexadecimal
0	00000	0
1	00001	1
2	00010	2
3	00011	3
4	00100	4
C	12	01100
D	13	01101
E	14	01110
F	15	01111
10	16	10000

10. An example of the sort of printout errors which can occur: first, both sync and phasing are lost; then sync is recovered, but phasing is still lost.

is relatively easy to construct very accurate and stable control circuitry, and as long as both the transmitter and receiver are driven and timed by highly stable master oscillators, they will remain synchronised. In addition, conventional methods, such as the use of phase-locked loop circuits, ensure that synchronisation is maintained even during frequency drift, as often occurs over a radio link.

Ensuring that the receiver scan starts at the same point as the transmit scan – normally the top of the page – is accomplished by the transmission of a phasing signal (*figure 9*) before the data transmission begins. At the receiver, the phasing signal is used to align a reference point in the middle of the blanking period of each line scan. Thereafter, the accurate frequency and phase control circuits, based on phase-locked loops, maintain synchronisation. *Figure 10* illustrates what happens if, first, both sync and phasing are lost, and second, sync is recovered without phasing.

The baseband resolution

We have mentioned earlier the importance of the resolution of a facsimile system, and it is now time to consider this parameter in greater detail. Resolution determines the amount of detail that the system can register. The limiting factor (apart from financial considerations) is the speed at which the system as a whole can react to changes at its input. The system, in this case, includes factors which are not actually in the signal chain, for example the size shape and quality of the scanning dot. In the receiver, similar factors in the reproduction mechanism play a similar role.

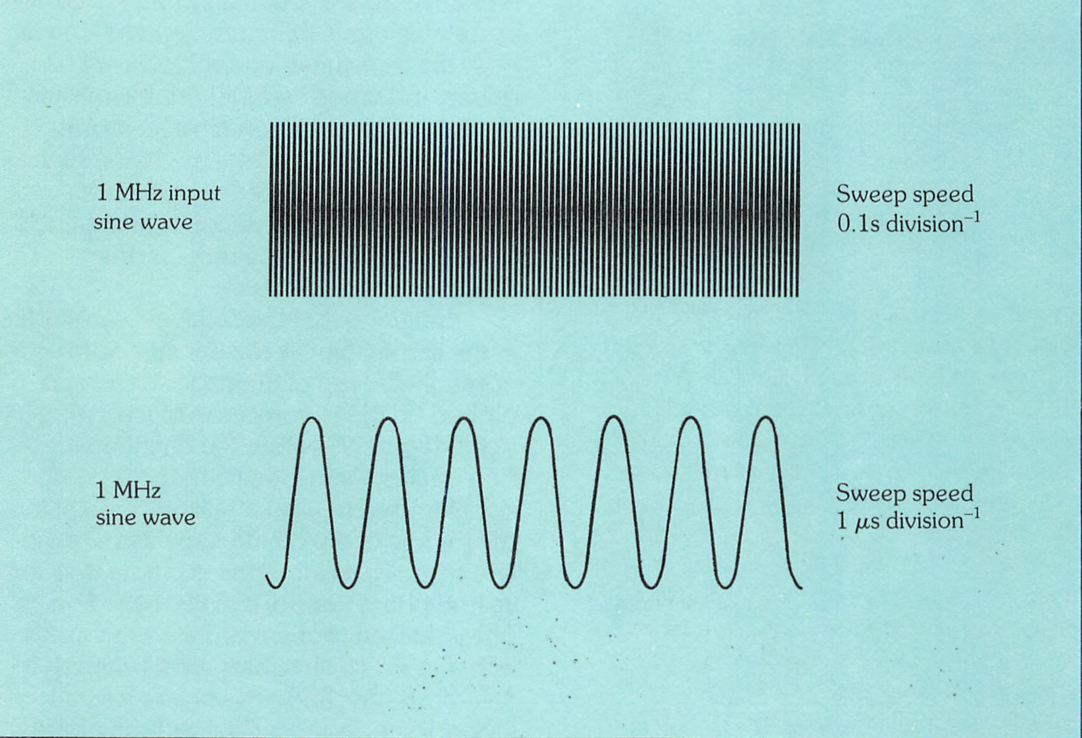
The input to a facsimile system is the image sensing device, and in this part of the system the first limiting factor is the speed at which the photodetector (photodiode, photomultiplier tube or linear image sensor) can react to changes in light intensity from the source document; a photodiode, for example, does not react instantly to the light falling on it. Photodetectors react very quickly, in human terms, with response times of about one microsecond, but nevertheless, the scanning mechanism must allow sufficient time for the sensor to react to the light, which means that the scanning speed must be matched to the response time of the detector.

To illustrate this, think what would happen if a slow photodiode with a response time of, say, one millisecond was coupled with a fast scan of one line per millisecond. This photodiode would not be able to keep up, and its output would not represent the reflected image from the page but would correspond to a mid-grey tone, representing the average of the reflected light during one line scan.

Scan speed is also an important factor in determining the resolution at the scanner. If, for example, you were to examine a sine wave with a high frequency (say, 1 MHz) on an oscilloscope with a long sweep time (say, 0.1s per division), you would see only a broad band of illumination across the centre of the screen. It would be impossible to resolve the peaks of the waveform, or estimate its frequency. However, if the oscilloscope sweep time is decreased (to say, 1 μs per division), the waveform becomes clearer and easier to resolve (*figure 11*).

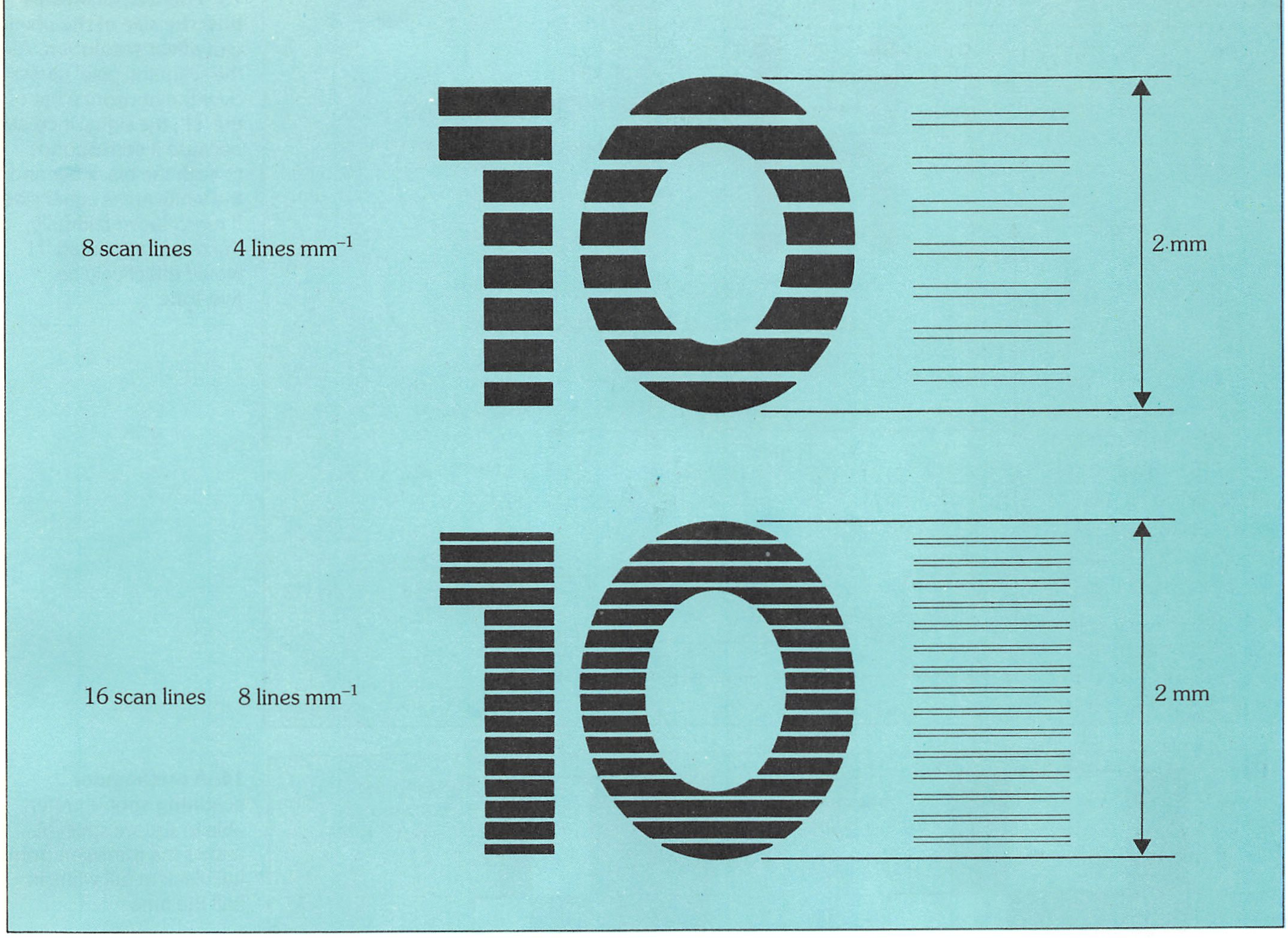
This same principle applies to facsimile scanners: the scanning speed must be adequate to resolve the most rapid tonal

11



11. The scan speed is important in determining the resolution. A sine wave with a high frequency observed on an oscilloscope with a long sweep time just looks like a broad band of light on the screen. Decreasing the sweep time, however, makes the waveform easier to see.

12



12. The line or scan density is governed by the size of the scanning spot. The finer the spot, the more scan lines and the better the definition.

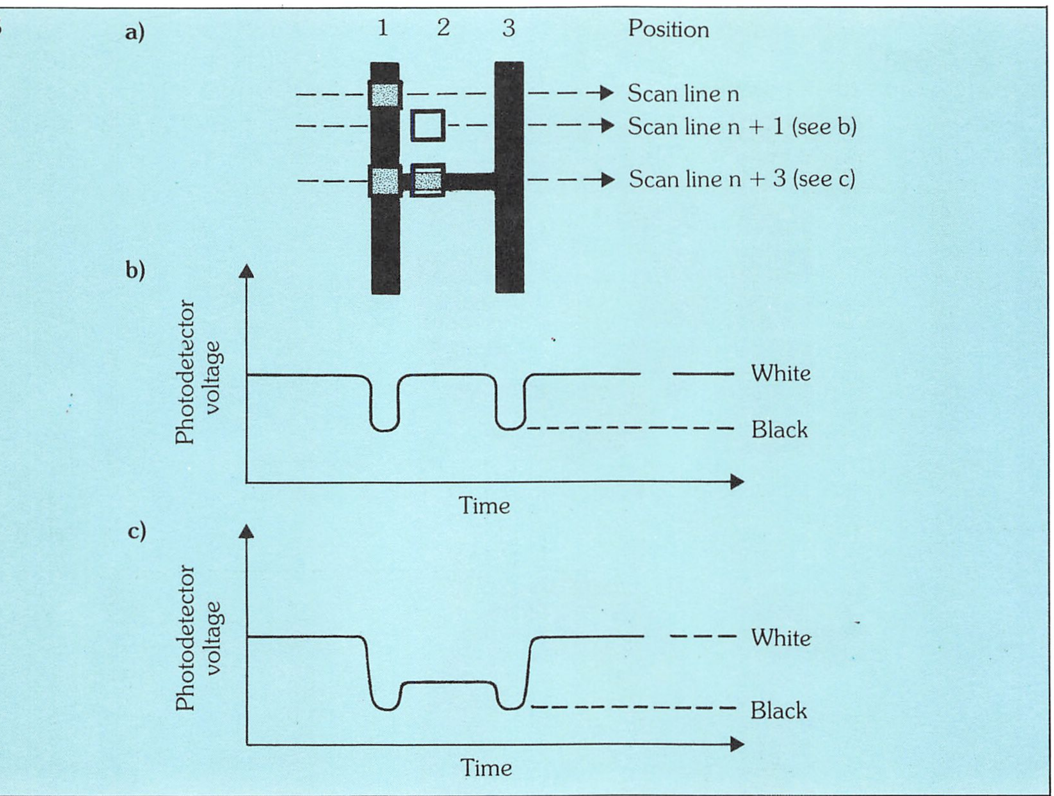
variations encountered in the document. Scan speeds used for document facsimile, for example, are typically 180 lines min^{-1} (3 per second) and, for higher resolution, 360 lines min^{-1} (6 per second).

The most important factor in defining resolution at the scanner, however, is the line or scan density. This is governed by the size of the scanning spot because, as *figure 12* shows, the finer the spot, the more scan lines and therefore the better the definition. In practice, the resolution of a facsimile system is normally specified by this scan density, in lines per millimetre. Typical densities in document facsimile are 3.85 lines mm^{-1} and 7.7 lines mm^{-1} for double resolution. As we shall see in a moment, much higher scan densities can be achieved and used for photo facsimile.

As we have said earlier, the scanning and sensing systems used in facsimile break down the image on the source document into picture elements or pixels, each of which is sensed by a photodetector device. Most scanning systems operate by sweeping a single small dot across a line, though in linear image sensor scanners the pixel is, in effect, stepped across the line in small increments equal to the width of the photosensing elements and at the rate of the clock pulses that shift the information out of the linear array.

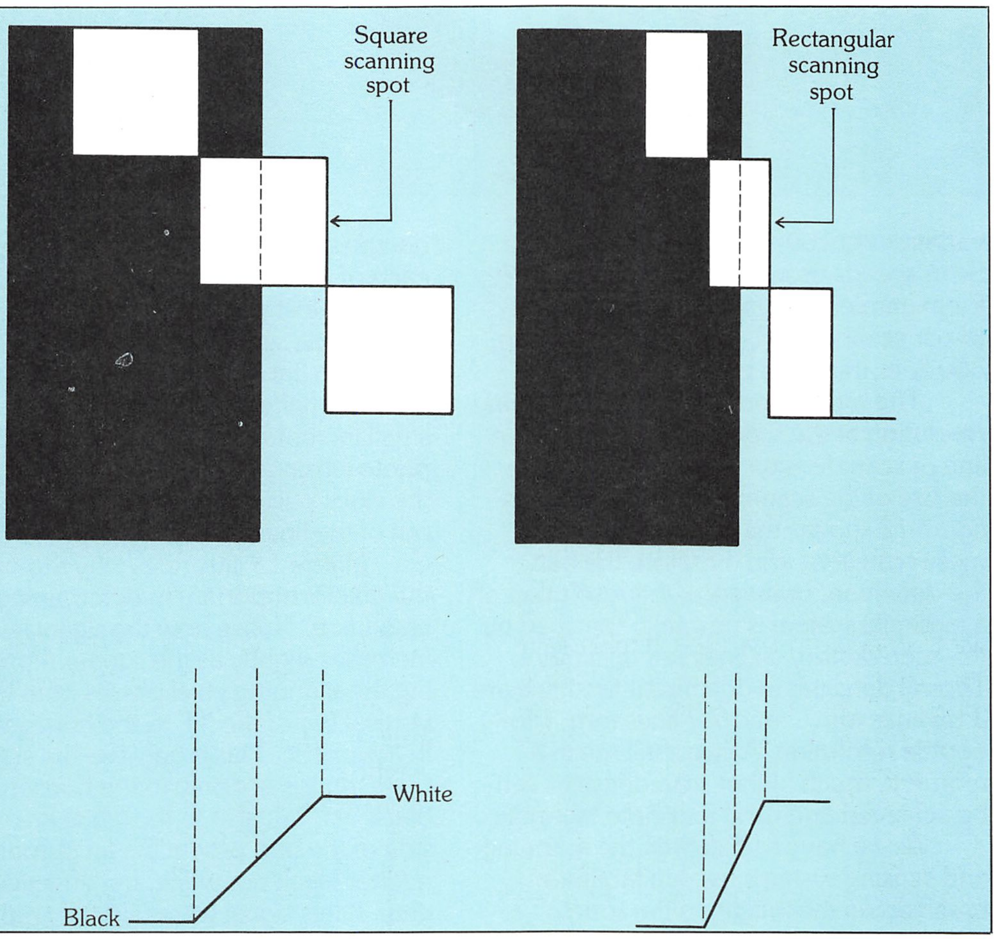
Figure 13 illustrates why the size of the pixel is important in determining the resolution. Notice how the signal level increases slightly as the square representing the scanning pixel passes from the vertical bar of the 'H' to the horizontal bar in *figure 13c*. This is because the signal corresponds not only to the horizontal black bar, but also to the white areas either side of the bar, producing an output that is neither black nor white, but effectively grey. If this signal were faithfully repro-

13



13. This demonstrates how the size of the pixel can affect resolution. As the scanning pixel passes over the horizontal bar of the 'H', the signal increases because it corresponds to both the black bar and the white areas either side. If reproduced faithfully, the resulting printed 'H' would effectively be two-tone.

14



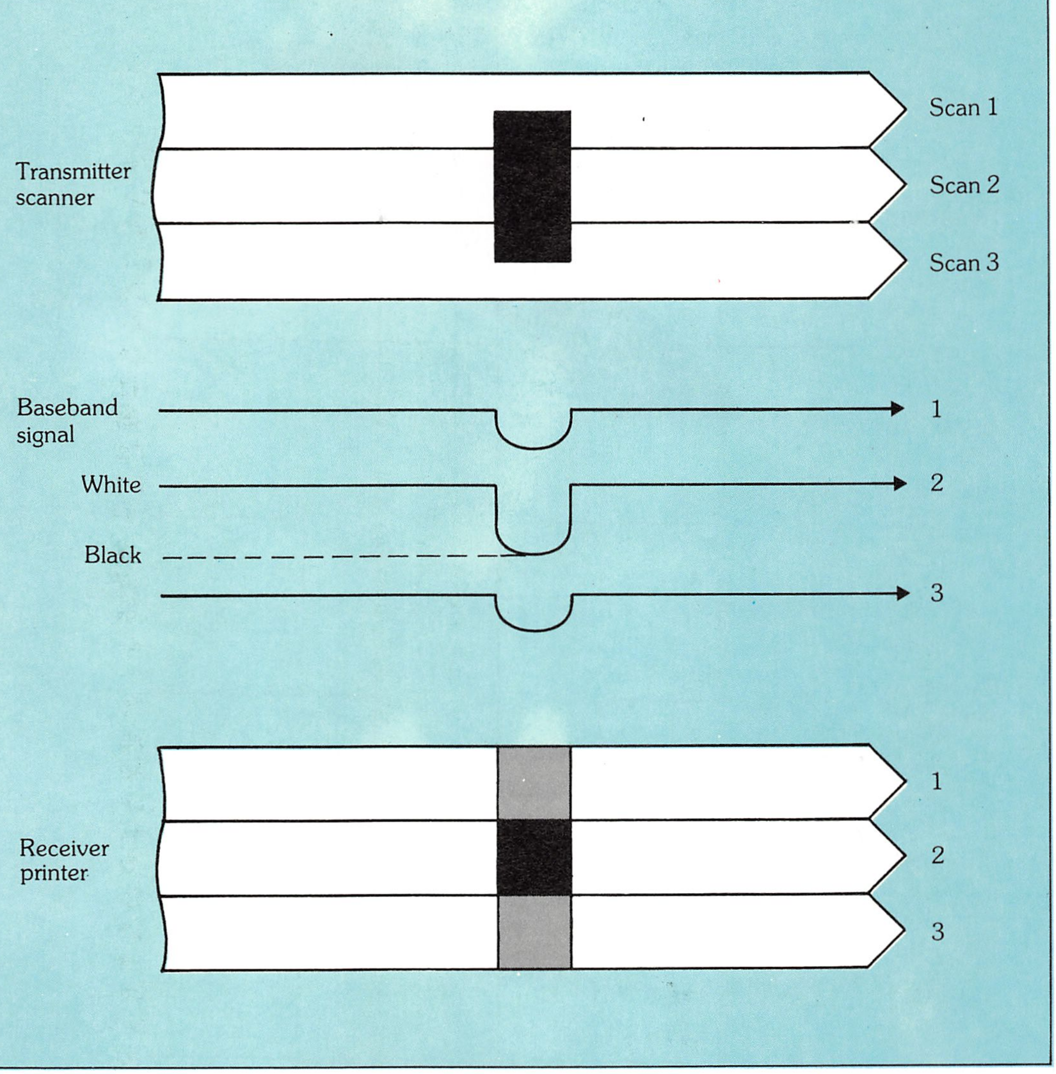
14. A rectangular scanning spot is preferable to square spot – it makes the transition from full black to full white in half the time.

duced, the printed 'H' would be two tone – black and grey.

In *Communications 5* we saw how a square scanning spot is preferred to a round one, because all parts of the source document are scanned. *Figure 14* shows that a rectangular scanning spot is even more preferable. The rectangle has the same height as the square but only half its width and, as you can see, it makes the transition from full black to full white in half the time. Remembering that in an analogue system, intermediate levels are reproduced as shades of grey, you can see that the faster the edge transition, the greater the tonal contrast in the reproduced image.

Earlier, we mentioned that the scanning density is one of the most important factors in determining resolution, a high scan density is also useful for another reason. It may take many scans to read one line of type and the space between adjacent lines. If one of the scans does not accurately align with the top (or bottom) of the printed line, there will be a blurring of the edges. This is shown in *figure 15,* where a line of type is covered by three scan lines (though it could be more, of course). When this happens, the vertical definition is blurred because the printing mechanism, as we shall see, can print only a whole trace, not part of one, and although the trace may be very small, it will still be larger than the black image area which was scanned at the same point. Therefore, the higher the scan density, the smaller the part-trace, and the better the

15



15. The printing mechanism can only print a whole trace. This example shows how three scan lines covering this image result in 'grey' areas top and bottom. Higher scan densities result in smaller part-traces and hence better resolution.

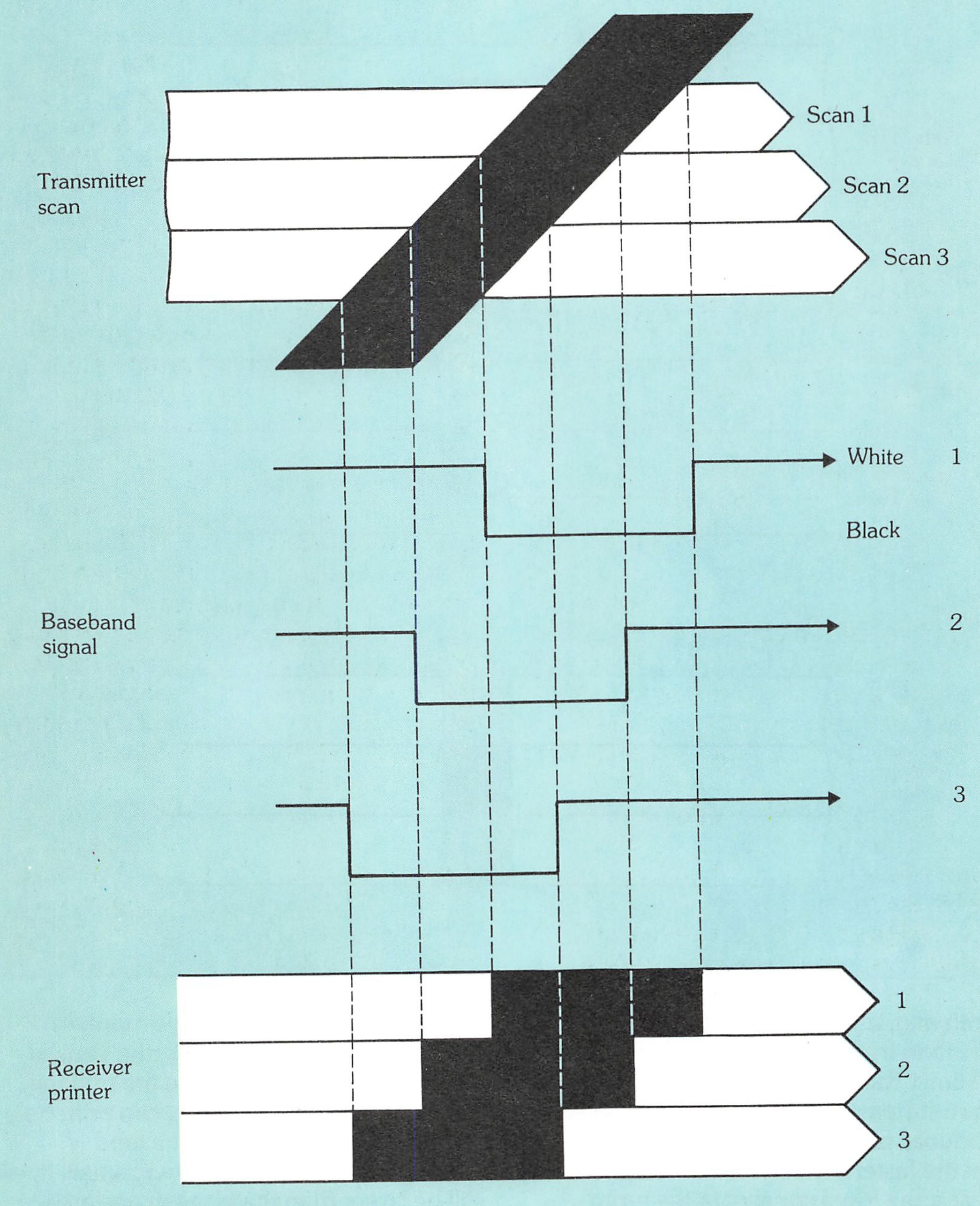
final resolution.

Figure 16 illustrates a similar problem that occurs when the source document is, for example, a line drawing running at an oblique angle to the direction of the line scan. You can see that the edges of the printed document are stepped or blurred.

Bandwidth limitations

We saw earlier that if the photodetector was slow to respond, more time had to be allowed to scan a line and, consequently, to transmit a full page of information; fast response detectors, on the other hand, can use faster scan speeds, resulting in shorter

16



16. Another example of how scan density affects the printed resolution.

transmission times. Fast scan rates, however, produce more information (that is, resolve greater detail) which requires a high quality, wide bandwidth communcations link.

But we know from the discussion on television transmission and reception, that unlimited bandwidth channels are not practical and so, as with all communications systems, bandwidth and transmission times are interrelated.

We can get a good idea of the time/bandwidth limitations in facsimile transmissions by comparing the information density of a variety of images and then looking at the rate at which that information can be transmitted. Taking a single frame of a television picture first, we know that there are 575 active lines in the 625 line U.K. television frame, and therefore the vertical resolution is 575 pixels. As the aspect ratio of the television CRT is 4:3, this gives us a total of approximately half a million pixels (see *Communications 4*). These pixels can either be illuminated (on) or not illuminated (off), so we can therefore consider each pixel to be a binary digit or bit, and so the total amount of information on the screen is approximately half a million bits, say, 0.5 Mbits. Compare this with the data content of an 8″ × 10″ fine grain photograph as shown in *table 1*.

Table 1
Common images and their data contents

Image	Data content
625 line PAL system television	0.5 Mbits
High resolution computer graphics screen	1.0 Mbits
Commodore 64 computer graphics screen	64 kbits
BBC computer Mode 4 screen	81.92 kbits
8″ × 10″ printed photograph in newspaper	200 kbits
8″ × 10″ printed photograph in glossy magazine	1.8 Mbits
8″ × 10″ fine grain photograph	150 Mbits
A4 printed page	1.6 Mbits

Remembering the bandwidth limitations on television transmissions, we can see that there are also going to be some limitations to facsimile transmissions. For example, 25 television frames each of 0.5 Mbits are transmitted each second, that is a data rate of 12.5 Mbits s^{-1} (25 × 0.5). But as each bit can either be on or off, the maximum *frequency* that can occur (when adjacent bits are in opposite states) is half of this number, i.e. 6.25 MHz. As a result, the bandwidth required to accommodate the maximum data rate of 12.5 Mbits s^{-1} is the same as the maximum, or **equivalent frequency,** that is 6.25 MHz.

Looking at this another way, we know that in the U.K. television system, the active portion of the line scan lasts for 52 μs, and so in this period 767 pixels (i.e. 575 × 4/3) must be transmitted. The data rate by this calculation is therefore 14.75 Mbits s^{-1} ($767/52 \times 10^{-6}$). (Remember, we're working in approximations here – the figures of 12.5 Mbits s^{-1} and 14.75 Mbits s^{-1} are close enough). The bandwidth required to accommodate the maximum data rate is therefore 7.4 MHz.

We could, of course, take longer to scan the television line. If the line scan period were lengthened, say to, 100μs, the bandwidth required would then be much less (7.66 Mbits s^{-1} or approximately 4 MHz). In this case, though, the picture would take longer to build up on the screen and the slower scan would result in decreased resolution. (A possible third alternative is to reduce the number of pixels in a line, but resolution would be lost here too). It is easy to see that the bandwidth and data rate result from a compromise between picture quality and speed.

Returning now to facsimile systems, we know that the usual scanning method for document fax is the linear image sensor. The horizontal resolution of such sensors is typically 1728 pixels, and if each pixel is regarded as a single bit, there are therefore 1728 bits. The standard scan rate is 180 lines per minute, and so the data rate of the baseband signal is 5182 bits s^{-1} (1728 × 180/60), or approximately 5.2 kbits; the bandwidth needed for this transmission is 2.6 kHz.

Again, the bandwidth can be calculated another way by defining the time in which the entire page has to be transmitted. For example, we know that the standard scan density corresponding to a scan rate of 180 lines min^{-1} is 4 lines mm^{-1}, so 1188 lines would be required to cover the full depth of an A4 (297 mm) page. However, as with television, not all lines are used and the number of active lines is actually 1144. Therefore the total

number of pixels or bits that the system must be able to resolve is 1,976,832 (1144 × 1728), or two million approximately.

Assuming that, for example, we want to transmit these two million bits in 6 minutes, we would therefore require a data rate of 5.6 kbits (2,000,000/360) and a bandwidth of 2.8 kHz.

Conveniently, both figures (2.6 kHz and 2.8 kHz) fall neatly within the bandwidth of the public telephone network. It is also possible, operating under a different standard, to transmit this page in 3 minutes, but at half the resolution – this is plainly an advantage where cost or speed, but not quality, is of prime importance.

Where both speed and resolution are important different methods must be used. For example, a firm of architects or an engineering company may need to transmit detailed plans, requiring high resolution, to the U.S. in less than three minutes. A second standard which may be used for this purpose calls for resolution of 2432 pixels across a page length of around 280 mm, and a scan density of 7.7 lines mm^{-1} But the data rate produced under this standard is far too high to be accommodated by normal voice-band radio or telephone links, so very powerful data and bandwidth compression techniques must be employed to reduce the effective bandwidth.

You can see, now, that amongst the many factors that influence the resolution of a system some are conflicting: the scan speed may need to be slow to match the response time of a particular photodetector, at the expense of resolution along the line and time needed for transmission; on the other hand, a quick scan, although resolving great detail and taking little time to transmit, requires a fast response photodetector and a high quality communications link. In practice, therefore, the resolution at the scanner requires a balance to be reached between the conflicting factors of scan speed, photodetector response time, the type of communications link, and the time that can reasonably be allowed to transmit a full page of information. Compromises must therefore be made to meet the requirements for a particular facsimile application – speed for weather fax, high resolution for newsfax, and so on.

Encoding

It was mentioned in *Communications 5* that one of the advantages of facsimile is its redundancy, or repetition of information, which means that small errors are negligible because the missing or incorrect information can be extracted or worked out from the remainder that was faithfully transmitted. This redundancy, however, is expensive in both time and bandwidth and various methods have been tried and tested for redundancy reduction or data compression. For example, **run length encoding** is particularly suited to document facsimile.

Run length encoding encodes at the source (i.e. it is **source encoding**) and it depends on the properties of the baseband signal. It relies on the fact that in a printed document or a black and white line drawing, there will be 'runs' of one colour only – lengths of time in which the signal is entirely white (such as when scanning between the lines of type) and other times when the signal is entirely black (such as when scanning along a solid line that runs across the page). Instead of sending every bit of the baseband signal during these periods, one digital word is sent by the encoder and this specifies the length of time for which the baseband signal is unchanged.

Table 2 lists the codewords for diffe-

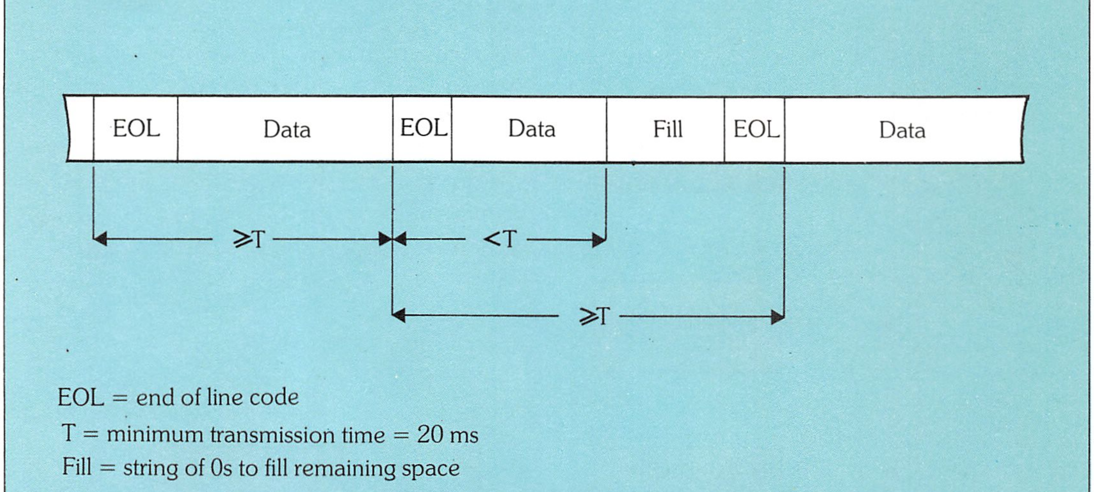
Table 2
Codewords for different run lengths of black and white

	Length	White code	Black code
Make-up codes	64	11011	000000111
	128	10010	0000110010000
	192	010111	000011001001
	256	0110111	000001011011
	320	00110110	000000110011
	1728	010011011	0000001100101
	End of line	000000000001	000000000001
Terminating codes	0	00110101	0000110111
	1	000111	010
	2	0111	11
	3	1000	10
	4	1011	011
	63	00110100	000001100111

Each run length = 1 terminating code or 1 make-up code + 1 terminating code

17. One-dimensional run length coding. Each line of data is followed by an end of line code. Lines shorter than 20 ms have a string of binary zeros incorporated as fillers to allow for the minimum transmission time necessary for the printing mechanism to catch up.

17



rent run lengths of white and black. The code used is a modified **Huffman code** (named after the mathematician who invented the method), with the shortest codewords assigned to the most frequent occurrences of white and black run lengths as typically found on a typed or printed document. Each word consists of a **terminating code,** which specifies run lengths from zero to 63, and possibly a **make-up code,** which specifies run lengths in multiples of 64. If a run length is less than 64 units, then only the terminating code is sent.

Run length encoding is extremely efficient as a codeword can be as short as two bits, and only as long as twelve. An entire line, therefore, could be specified by just two bits, rather than the 1728 it would otherwise take.

However, the code for a line can be too short, and transmitted too quickly. Remember, time must be allowed for the printing mechanism to catch up, and so a minimum run length of 20 ms is specified for each line – any lines shorter than this must have a string of binary zeros incorporated, as fillers. In addition, each line of data is followed by an **end of line** code that instructs the printer to start a new line scan (*figure 17*).

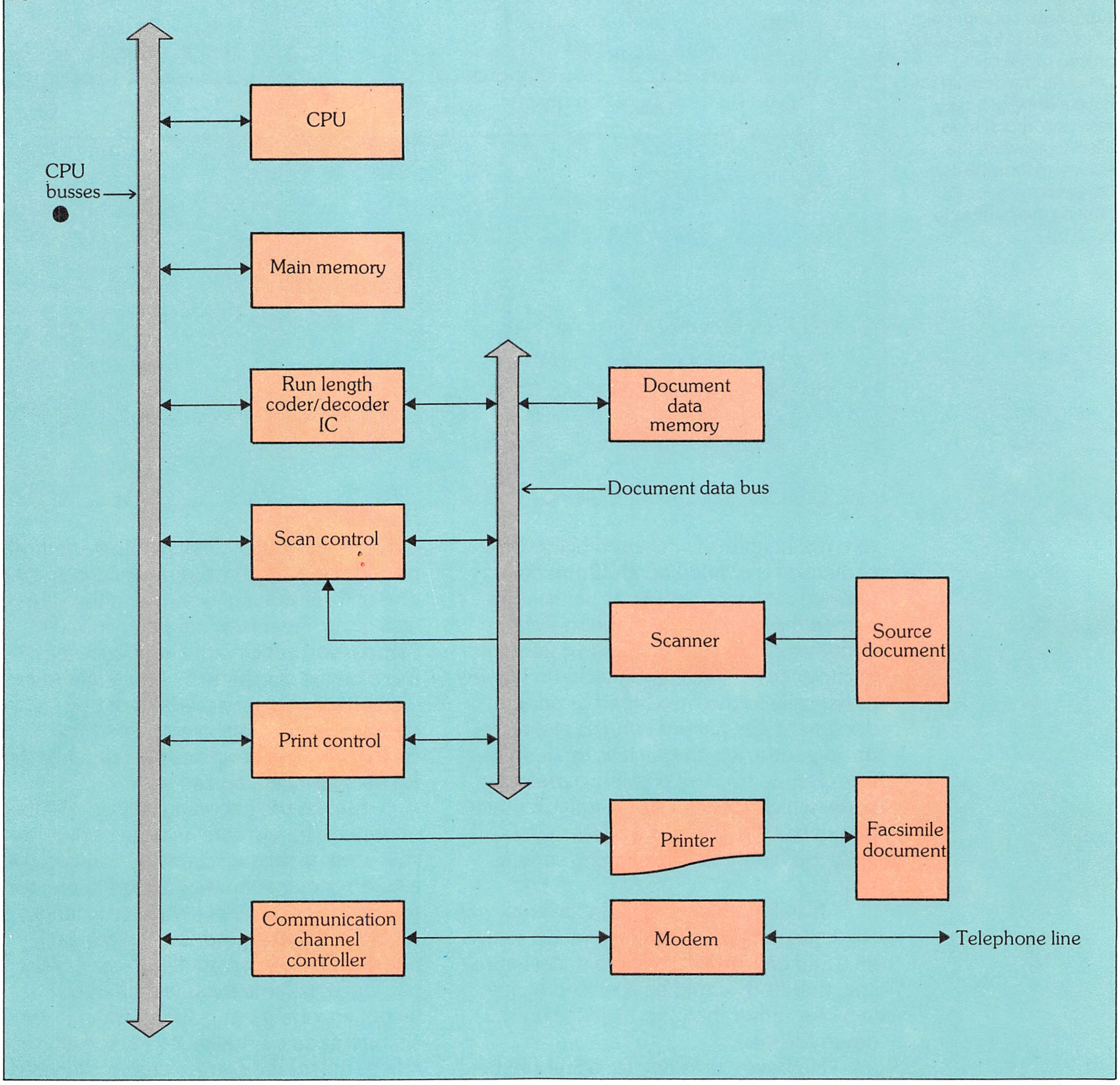
The method previously described is known as **one-dimensional run length coding,** because it operates on a one-dimensional line of data. Another method, **two-dimensional coding,** selects between one of three coding formulae at the start of each scan line. Here, data is encoded by comparison either with a reference element on the same line or with a reference element on the previous line, or one-dimensional coding is selected. Once coded, that line becomes the reference line for the next scan, and so on.

Run length encoding is not a real time operation. Image data, usually from a linear image sensor scanner is converted to a binary stream and then stored in memory so that the encoder can detect run lengths. A codeword is then substituted for each run length, and transmitted instead of the bits that constitute the raw data. Run length encoding using a single integrated circuit (made by Advanced Micro Devices) that performs the complicated coding and decoding operations is shown in *figure 18.*

Continuous tones

As we have seen, the analogue baseband signals for continuous tone photographs are produced by scanning a spot or rectangle in a smooth sweep across the document – they are then converted into digital signals by an ADC – this method produces the best resolution. However, a continuous tone print contains a considerable amount of information, and so needs to be scanned at maximum line density.

18



The best laser scanners are now capable of producing densities of up to 32 lines mm^{-1}, and this figure will be improved upon in the future.

If we assume that the resolution across a scan line is the same as the resolution down the page, then for a 200 mm × 250 mm (8″ × 10″) photograph, a modern system may resolve: 51.2 million pixels (32 × 200 × 32 × 250). Each pixel is then converted into a digital word of 8 bits (each 8-bit word is capable of specifying 256 different shades of grey), so the volume of data produced is actually multiplied by eight, giving a total of about 410 million bits.

Obviously, this amount of data could not be transmitted over a standard voice band link without very efficient compression. Run length encoding, though very efficient for black and white documents, is not useful for continuous tone transmissions simply because runs of black or white do not often occur. Instead, **channel**

18. The role of the run length coder-decoder IC in a complete facsimile system.

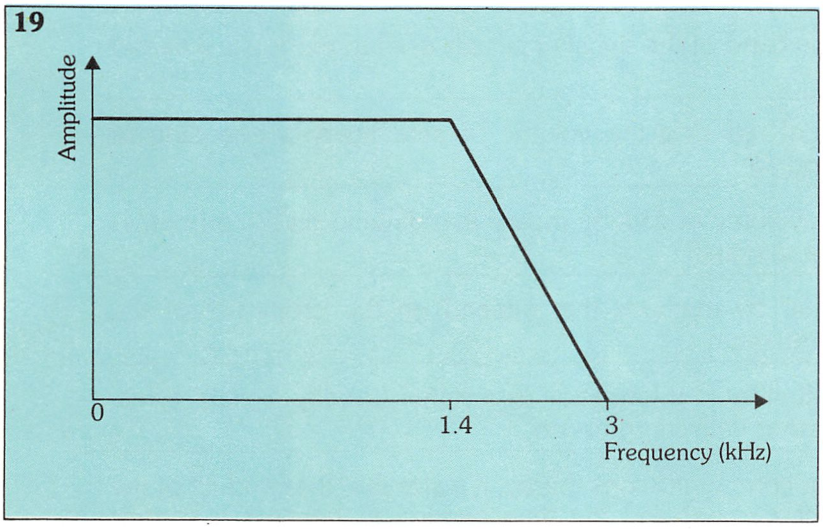
encoding methods, borrowed from the field of data communication, are used to process the raw data into a form that does not require extravagant bandwidths: data compression ratios of between 3:1 and 90:1 can be achieved, depending on the material being sent, and on average, the compression ratio is around 15:1.

Other tricks, however, must still be used, the most common being to multiplex the output signal onto *two* telephone lines, thus achieving an effective transmission bandwidth of nearly 6 kHz. These methods will be discussed in greater detail in subsequent *Communications* articles.

Modulation

In facsimile transmission, commonly used modulation methods are amplitude modulation (AM), frequency modulation (FM), and, for digitally coded signals, pulse code modulation (PCM). Different methods are selected for different applications, usually to meet the bandwidth requirements of the communications channel, but usually a variant of one of the standard techniques is chosen to give some degree of bandwidth compression. For example, signal sideband AM is used rather than double sideband AM, and different varieties of PCM are also used to compress digital data.

19. The modulation method most frequently used in facsimile transmission is vestigial sideband AM. This graph shows the characteristics of the sideband signal.



Such channel encoding is necessary because although the bandwidth required for a basic resolution transmission (2.6 kHz or thereabouts) is within the nominal bandwidth of voice frequency communications channels, it is a little tight. Signal components with frequencies near the top and bottom of the range, for example, may be lost, with a consequent reduction in resolution at the receiver.

We have already looked at channel encoding methods – otherwise known as **modulation methods** – that require less bandwidth than others. In fact, the method most frequently used is vestigial sideband AM, the same method used for the transmission of a television composite video signal that we saw in *Communications 4*. *Figure 19* shows the characteristics of the vestigial sideband signal for facsimile transmission.

Another approach frequently used when the link is a radio channel, is **frequency shift keying,** in which two different audio tones are transmitted, one of 2100 Hz for black and another of 1300 Hz for white. There is no carrier frequency in this system, only tones corresponding to either black or white.

Standards

One facsimile machine can only talk to another if they share certain characteristics: for example, they must both scan lines at approximately the same rate, and at about the same density, if resolution is to be preserved. It is therefore necessary to have internationally accepted standards of facsimile equipment and signals to ensure that communication between different makes of equipment is possible.

The most important organisation with respect to facsimile standards is CCITT (International Telegraph and Telephone Consultative Committee) and all of thestandard figures that have been mentioned in this description of facsimile equipment are CCITT recommendations. There are different standards for different applications, of course, but provided they are maintained, a facsimile machine of one standard can always talk to a machine of another standard, to some degree. Even so, there remain problems, not least because CCITT is not the only body to specify standards for facsimile communications. Other standards, more or less corresponding to the CCITT standards, are set by FCC (U.S. Federal Communications Commission), WMO (World Meteoro-

logical Office), CCIR (International Radio Consultative Committee), various military organisations, and others!

The degree to which different types of facsimile equipment are compatible is measured by the **index of co-operation,** and distortion may result when the indices of the transmitting and receiving equipment are too far out of range.

The problem of different standards has been solved, in part, by the latest generation of facsimile equipment. Very sophisticated facilities are now provided which allow the transmitter to communicate directly with the receiver, without the aid of a human operator. The transmitter interrogates the receiving machine to determine which standards it meets, and can then adjust its own operating system to match. Facsimile transmission and reception can therefore be totally automatic. Source documents can be stacked in a tray and left for the night, and when telephone calls are cheapest the transmitter automatically dials the required number, adapts itself to suit the operating system of the receiving equipment, and proceeds to transmit the documents. And with the encoding techniques we have looked at, they can do that at a rate of one A4 page every 15 seconds!

With advances such as these, facsimile has progressed considerably since Bain's pendulum. In the future, the capabilities of facsimile should be still more impressive.

Glossary

pulse amplitude modulation	a pulse modulation scheme in which the amplitude of the pulses are made to vary in proportion to the instantaneous amplitude of the modulating waveform
baseband	the frequency band of the signal or signals used to modulate a carrier frequency
data compression	the process of reducing the volume of data by data encoding or channel encoding
data encoding	reducing the volume of data by means that depend on the properties of the baseband signal
channel encoding	data reduction by methods that depend on the properties of the transmission medium
redundancy reduction	eliminating duplicated information in a signal (usually digital). Effectively the same as data compression
Hartley's law	a basic law in communications theory, it states that the product of the bandwidth of a signal and the time needed to transmit a specific amount of information is a constant

DIGITAL & SOLID STATE APPENDICES

Frequency, time and event counters

Counters

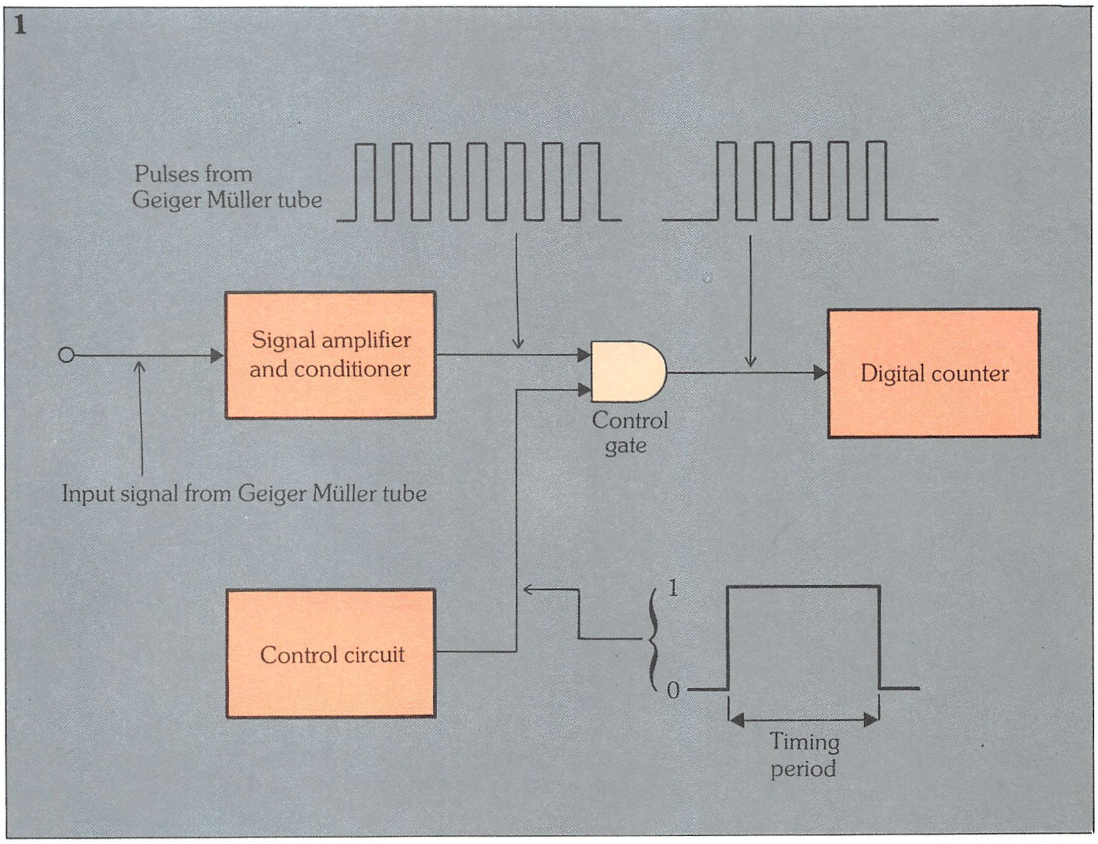
Frequency, time and event measurements are undertaken by test equipment given the generic name **counters.** Counters are generally digital devices which display the corresponding measurement on a digital display of some form. Together with multimeters, signal sources and oscilloscopes they form the central core of general purpose test equipment.

Various methods are used by such counters to measure the parameters of frequency, time and number of events, but nearly all these methods depend on a comparison between the signal being measured and a reference signal. As we shall see, there are a number of ways in which this can be achieved.

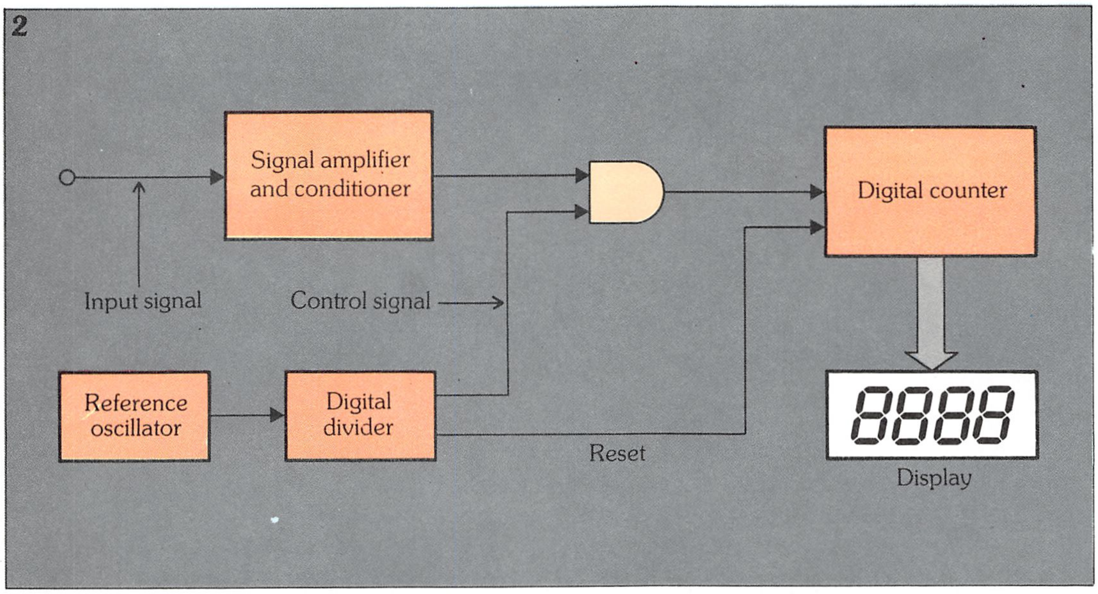
Event counting

It is often necessary in electronics to count a number of events: the classic example being the pulses generated by a Geiger Müller tube, the number of which over a period of time provides an indication of radioactivity level.

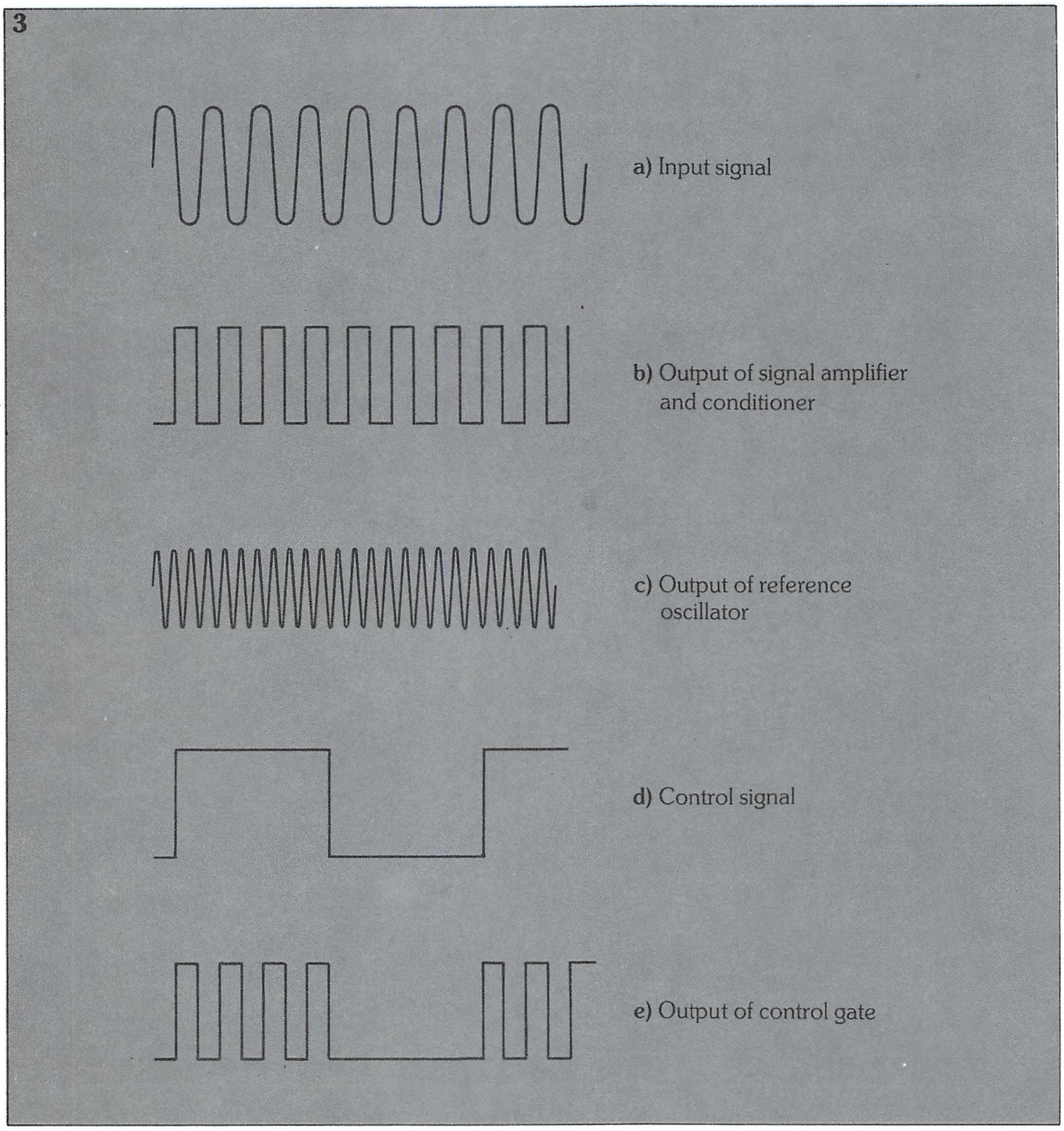
Figure 1 shows a block diagram of a simple mechanism to count events, which uses a digital gate controlled by the period of time over which the measurement is to be taken. During the time period, the control input to the gate is logic 1 and therefore all incoming pulses, after amplification and signal conditioning, are counted by the digital counter. At the end of the time period, the control input is logic 0 and no further pulses may be counted. This principle of gating a signal is the basis



1. Block diagram of a simple mechanism to count events. The input to the gate is controlled by a timing period produced by a clock.

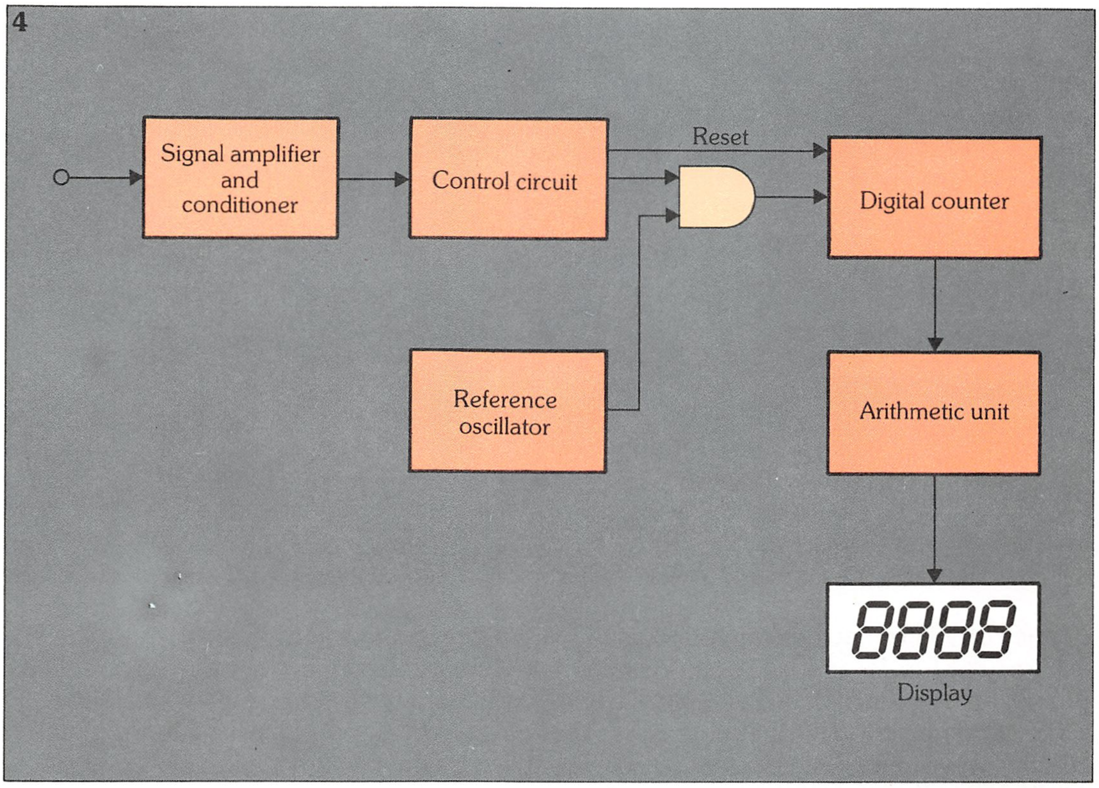


2. A direct gated counter can display the frequency of an input signal. The output from a reference oscillator is divided by a digital circuit to give the required timing period to control the gate.



3. Possible timing diagram of the type of direct gated counter illustrated in *figure 2.*

4. A reciprocating gated counter. This is an adaptation of the direct gated counter, in which the input signal is used to gate the reference oscillator signal.



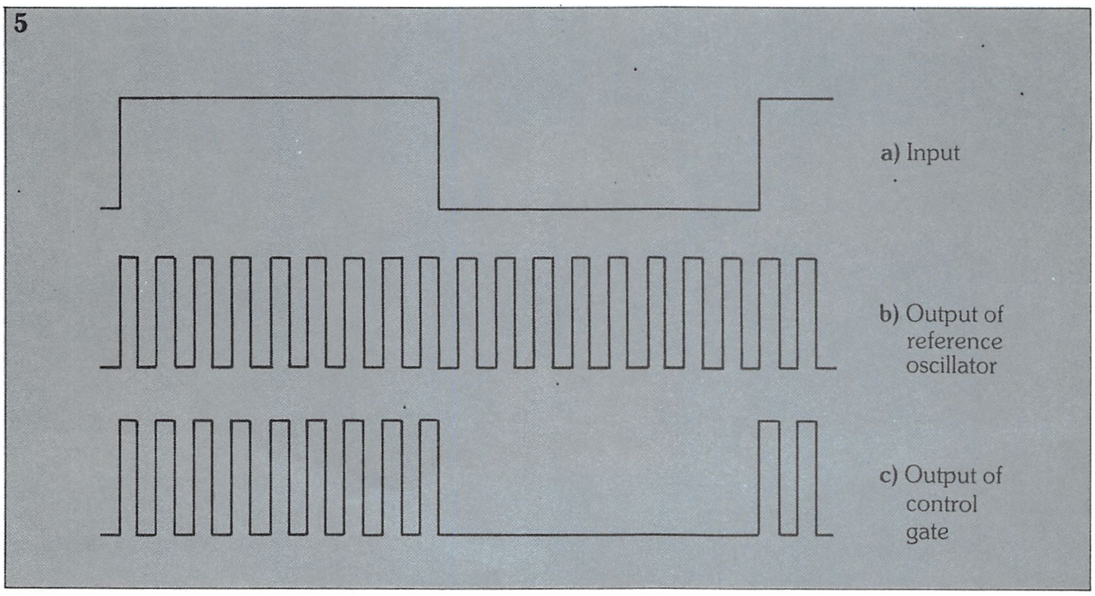
upon which all counters are built.

In the example of *figure 1* the number of pulses generated by the Geiger Müller tube may not be large (say, 10 per minute) so, in this case, the measurement takes place over a long period of time – a few minutes (or hours perhaps). Under such conditions the user will normally control the gate with a manually operated switch to provide the control signal. If the user miscalculates the timing period by a second or so, the measurement will not be too inaccurate.

If events occur at a much higher rate, on the other hand, the timing period during which the gate is open is much more critical. If, say, a regular train of pulses occurs at a rate of 1000 min^{-1}, and these pulses are counted for 61 seconds, a display of 1016 would be obtained. If it is wrongly assumed that the timing period is exactly one minute then an inaccurate measurement is the result. This one second error, however, is not as significant if the timing period was, say, ten minutes and not one minute.

To generate accurate timing periods, counters generally use a stable high frequency reference oscillator, the output of which is divided by a digital circuit to give the required timing period to control the gate. This method produces a **direct gated counter,** shown in *figure 2,* which may be used to display an input signal's frequency. For example, say a signal of 999 Hz is applied to the counter of *figure 2*. The reference oscillator has a frequency of, say, 32,768 Hz and the digital divider has a division ratio of 1/65,536 (i.e. 2^{-16}). In this example, the control gate is thus open for one second in every two seconds. During the second the gate is open, the digital counter counts each cycle of the input signal. During the following second, this count (which totals 999) is displayed. Immediately prior to the next second (another count period), the digital counter must be reset, of course, ready to commence the next count from zero.

Figure 3 shows a possible timing diagram of a direct gated counter in the style of *figure 2,* in which an input signal is counted and displayed as a frequency. Obviously, different input signal frequencies may be counted and displayed by this method, using different digital divider division ratios. The lower limit to frequency of applied signals depends largely on the length of time the user is prepared to wait for the frequency to be counted (a fre-



5. Timing diagram which can be obtained from a reciprocating gated counter.

quency of 10 Hz may be counted to an accuracy of 1 in 10, in a one second timing period – a frequency of 1 Hz requires a 10 second timing period for the same accuracy). The upper limit to frequency of applied signals depends on the speed of the logic circuits used in the digital counter. For example, direct gated counters which may display frequencies of over 500 MHz are available.

Measuring lower frequencies

As we have seen, the direct gated counter cannot be used to measure and display low frequencies directly. The principle of gating a digital counter, however, can. *Figure 4* shows an adaptation of the direct gated counter in which *the input signal is used to gate the reference oscillator signal.* This, of course, is the *exact opposite* of the direct gated counter, in which the reference oscillator signal gates the input signal. *Figure 5* shows a timing diagram which may be obtained in such a **reciprocating gated counter**.

The counted signal in the reciprocating gated counter must be inverted by an arithmetic unit before display, to allow the input signal frequency to be displayed.

Low frequency multiplying counter

Figure 6 illustrates another method of measuring low frequency signals, in which a phase-locked loop circuit is used to **multiply** the frequency of the input signal before gating, counting, and display using the direct gated counter technique. The output frequency of the voltage controlled oscillator within the phase-locked loop is equal to the input signal frequency multiplied by the divider circuit's division ratio.

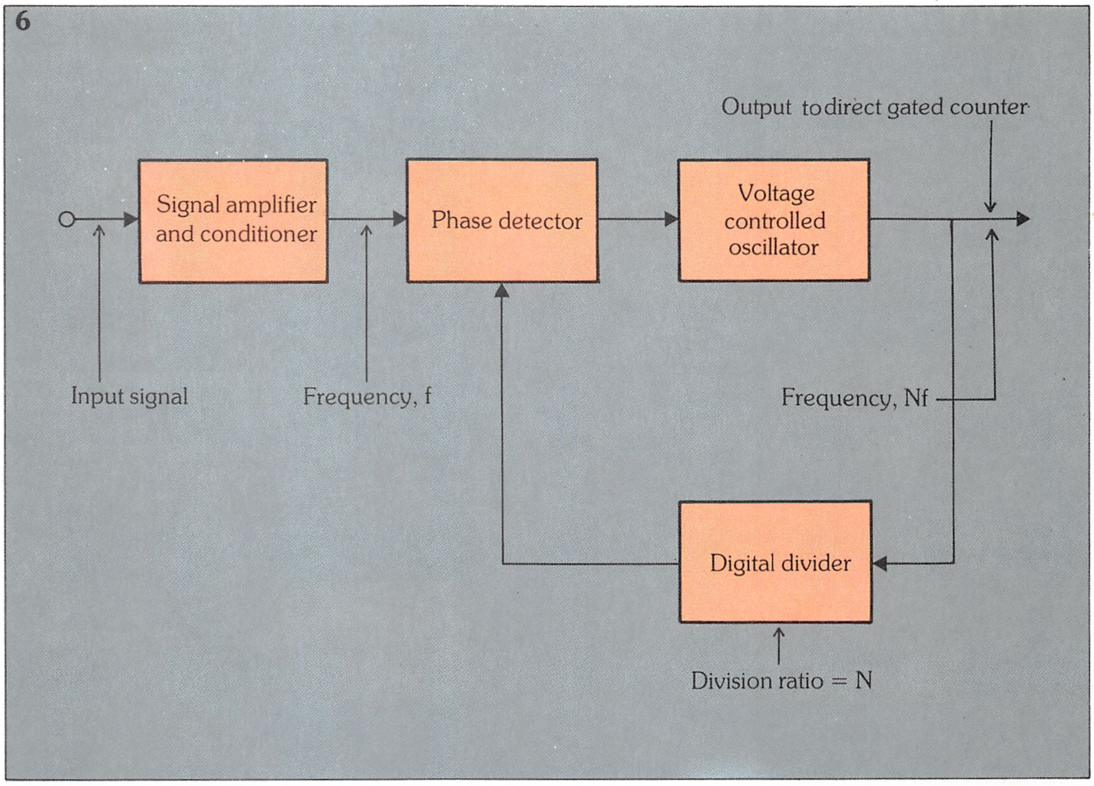
Measuring higher frequency signals

Although we said earlier that the upper limit of frequency of applied signals to a direct gated counter depends on the speed of the logic circuits in the digital counter, it is possible to add extra logic circuits to extend the upper limit. Such an extra circuit is termed a **prescaler** and is shown in *figure 7.* The use of a prescaler is possible because it is purely a dividing circuit (in the case of *figure 7,* the prescaler divides the incoming signal frequency by 8) and is inherently faster in operation than a digital counter. To enable the digital counter to count and display the correct frequency, the timing period must, consequently, be longer (by the same factor as the prescaler – in the example of *figure 7,* by 8 times). Input signal frequencies over about 2 GHz may be measured using a prescaler type frequency counter.

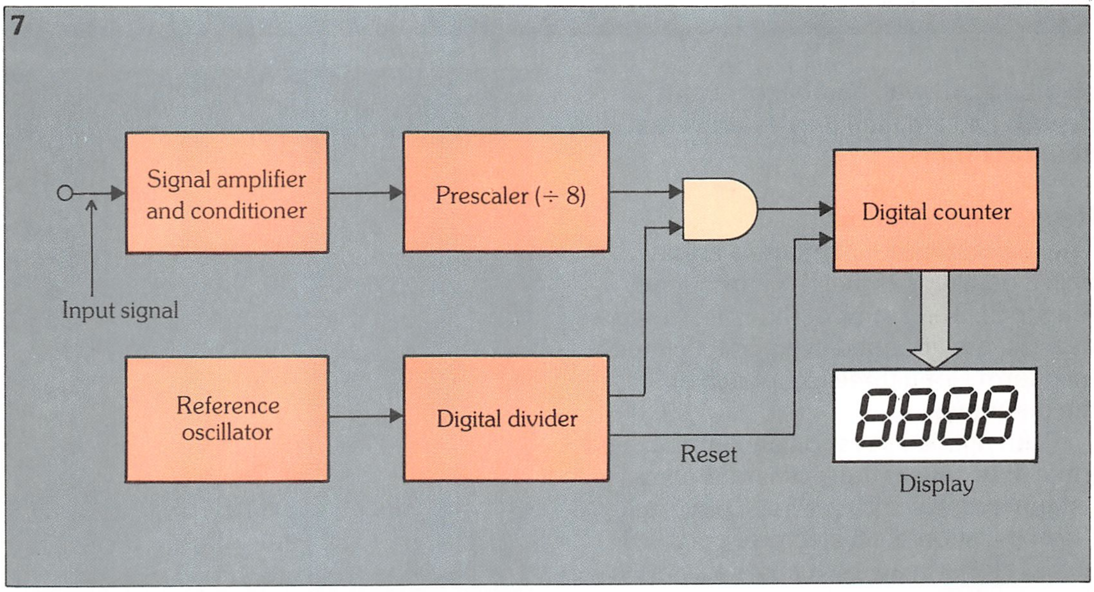
Heterodyne conversion

Even higher frequency input signals may be measured (over 20 GHz) using a technique used in radio communications called **heterodyning** (see *Communications 4*). Heterodyning, or **beating** as it is sometimes known, is used in such a counter to generate a signal whose fre-

6. A phase-locked loop circuit can be used to multiply the frequency of the input signal before the gating, counting and display operations. This is another method of measuring low frequency signals.



7. With a prescaler circuit dividing the frequency of incoming signals, a counter can be used to measure higher frequencies.

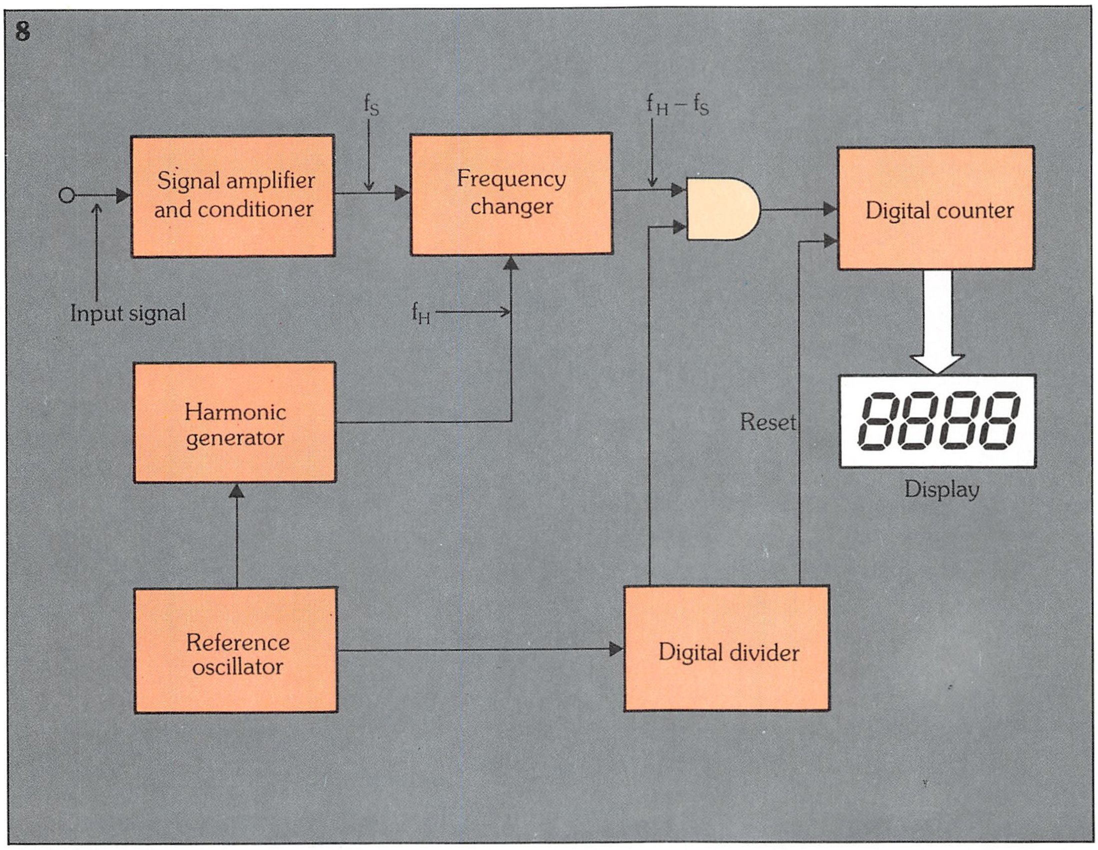


quency is the *difference* between the input signal frequency and that of a known harmonic of a reference oscillator. As this technique produces a signal of *lower* frequency than the input signal frequency, the process is known as **heterodyne down-conversion,** and a counter using the process is shown in *figure 8*.

The heterodyne signal is then counted and displayed by a normal direct gated counter. Input signal frequency is calculated by the user as the displayed count *plus* the harmonic frequency.

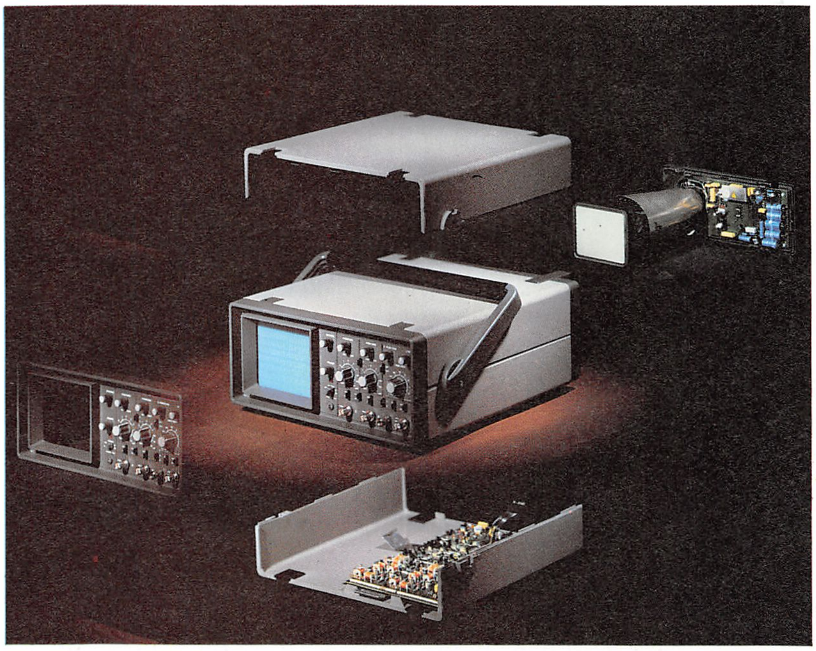
Counter types

The examples we have seen here of counters, are used to measure and display the frequency of an applied input signal. If a counter is designed to do solely this task, it is termed a **frequency counter.** Generally, however, counters are designed to measure and display other parameters, too



8. A counter using the heterodyne down conversion process.

Below: 'exploded' photograph showing the relative positions of CRT and circuitry inside an oscilloscope. This oscilloscope can trace frequencies up to 15 MHz.



(e.g. time interval, number of events, period) and are then termed **universal counter timers**.

Reference oscillators

The majority of counters use a crystal based oscillator as their high frequency reference. The atoms of the crystal used – quartz – are arranged in a perfectly repeated form along three axes, each at 90° to the other two.

An applied stress on any one of these three axes causes the atoms of that axis to generate an electric current. Conversely, the application of an electrical potential across atoms in an axis causes a stress in the crystal. The crystal thus acts like a highly selective parallel resonant circuit and may thus be used as the basis of an **internal reference oscillator** whose frequency is extremely stable.

External reference oscillators

Although the accuracy and stability of quartz crystal internal reference oscillators are good enough for the majority of measurements, some applications exist in which even greater accuracy and stability is demanded. For this reason, many counters are designed to accept a signal from an **external reference oscillator.** Such a signal may be derived from, say, a broadcast radio transmission of a highly accurate and stable national standard reference frequency.

Glossary

counters — generic name to describe test equipment used to measure and display frequency, time, period, number of events etc.

direct gated counter — basic counter principle in which an internal reference oscillator is used to gate the applied input signal before it is counted and displayed

external reference oscillator — reference oscillator used with a counter, allowing measurements of higher accuracy than would be possible with an internal reference oscillator

frequency counter — a counter used specifically to measure and display the applied input signal's frequency

heterodyne down-conversion — counter principle in which the high frequency input signal is heterodyned with a known harmonic of the reference oscillator to produce a lower frequency signal (which is the difference between the harmonic frequency and the input signal frequency)

internal reference oscillator — reference oscillator, based generally on the quartz crystal principle, within a counter

low frequency multiplying counter — a counter using the phase-locked loop principle to multiply the applied low frequency input signal by the digital divider's division ratio. This multiplied frequency may then be measured and displayed using a direct gated counter

prescaler — device used with a direct gated counter to increase the upper limit of applied signal frequency by dividing the applied signal

reciprocating gated counter — counter principle used in low frequency measurements, in which the input signal is used to gate the signal from the reference oscillator. The counted result is inverted before display

universal counter timer — a counter capable of measuring and displaying frequency, period, time interval, number of events etc.

10

DIGITAL & SOLID STATE APPENDICES

Spectrum analysers

Signals in the frequency domain

In all of the CRT-based test equipment we have looked at so far, oscilloscopes (basic and storage) and logic analysers, input signals have been displayed as amplitudes with reference to time. Such displays are said to be in the **time domain**, and may be used to form a graph of voltage against time on the CRT screen.

There is another way of displaying signals, in which a graph is formed of voltage against frequency. We have, in fact, seen examples of these **frequency domain** graphs in Bode plots and in frequency spectrums of transmitted radio and television signals. A CRT-based item of test equipment used to display a frequency domain graph of an input signal is known as a **spectrum analyser**.

Figure 1 illustrates the basic difference between domain and frequency domain displays. *Figure 1a* is a three dimensional representation of a possible input signal which consists of a number of sine wave components. Sine wave 1 (shown in green) is the lowest frequency component, with a frequency of f and an amplitude of 0.81. Sine wave 2 (in red) is at a frequency of 3f (i.e. the third harmonic of sine wave 1) and has an amplitude of 0.09. A third sine wave, sine wave 3 (in blue) has a frequency of 5f (the fifth harmonic) and an amplitude of 0.03.

Time domain analysis of these three sine wave components leads to the graph, which may be displayed on an oscilloscope screen, shown in *figure 1b*. From an earlier *Basic Theory Refresher* article on Fourier series we know that addition of three sine wave components of these frequencies and amplitudes, that is:

$$f(x) = 0.81 \sin \omega t - 0.09 \sin 3\omega t + 0.03 \sin 5\omega t$$

gives a triangular wave with an amplitude of 1. It is, in fact, a triangular wave which we see in time domain analysis with the use of an oscilloscope.

Frequency domain analysis of the three sine wave components produces a graph as shown in *figure 1c*, where the amplitudes of the components are shown in relation to their frequencies. Such a graph may be displayed on the screen of a spectrum analyser.

Displays of frequency domain analysed signals may often be more useful than time domain analysed signals, because the individual representation of components within the signals may provide a greater understanding. For example, extra harmonics, producing distortion to a signal, can be easily detected and traced using a spectrum analyser. An oscilloscope display of the same signal will probably not allow a useful analysis.

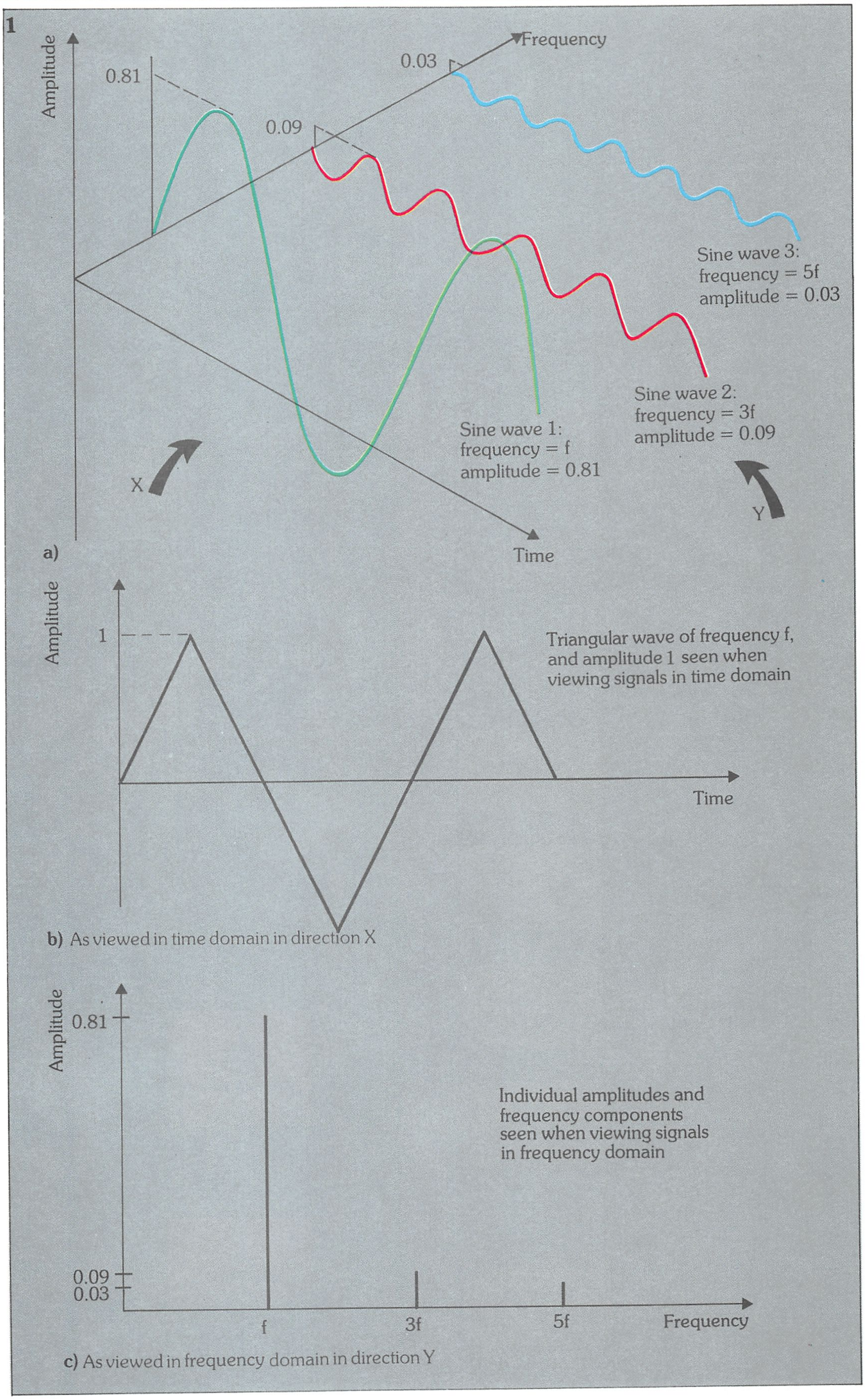
Displaying frequency domain

An ideal spectrum analyser must be capable of measuring and displaying signal components at individual frequencies within the range of frequencies under observation. Generally, however, it is not possible to measure and display individual components, so displays of groups of components within the range are used. There are three basic methods used in spectrum analysers to produce the required frequency domain display.

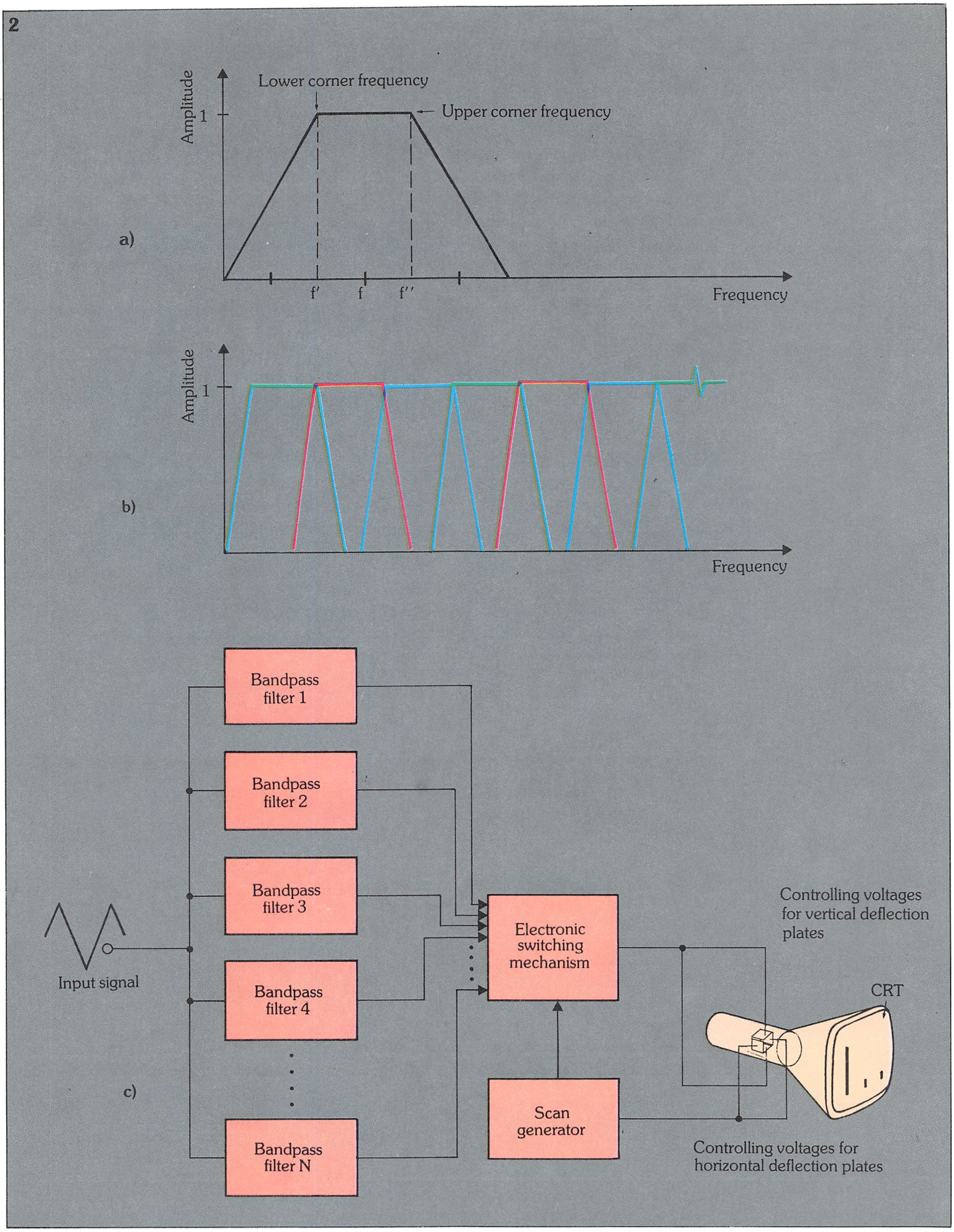
Real-time analysis

A **real-time spectrum analyser** contains a number of **bandpass filters,** each of which has a frequency response as shown in *figure 2a*. Each bandpass filter may be passive, and thus have a maximum gain of 1(0 dB) as in *figure 2a,* or active with a greater gain. By aligning the responses of the filters such that the upper corner frequency of each filter intersects with the

1.(a) Three-dimensional representation of an input signal consisting of a number of sine wave components.
(b) Graph resulting from time domain analysis, which can be displayed on an oscilloscope screen.
(c) Graph produced by frequency domain analysis, as displayed on the screen of a spectrum analyser.

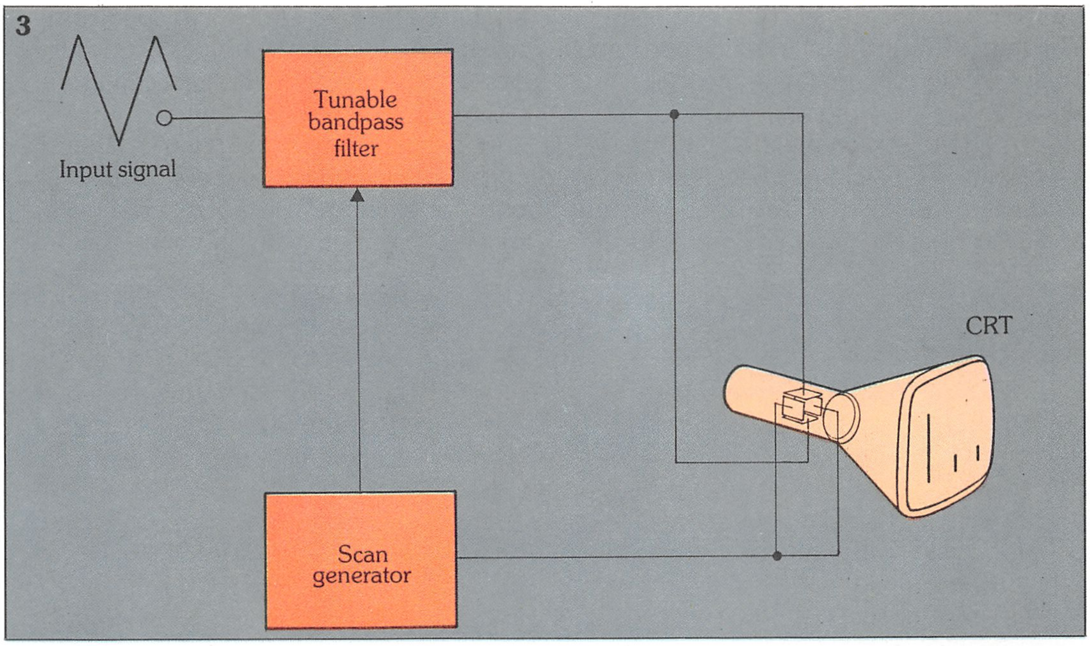


2



Amplitude
1
Lower corner frequency
Upper corner frequency
f′
f
f″
Frequency
a)
Amplitude
1
Frequency
b)
Bandpass filter 1
Bandpass filter 2
Bandpass filter 3
Bandpass filter 4
Bandpass filter N
Input signal
Electronic switching mechanism
Scan generator
Controlling voltages for vertical deflection plates
CRT
Controlling voltages for horizontal deflection plates
c)

2.(a) Spectrum of a single filter of a real-time spectrum analyser; **(b)** showing how filters are aligned; **(c)** block diagram of a real-time spectrum analyser.



3. Swept tuned spectrum analyser. A single bandpass filter, tunable over a range of frequencies, is used.

lower corner frequency of the adjacent filter, as in *figure 2b,* the real-time spectrum analyser of *figure 2c* may be built up.

The bandwidth of each filter defines the frequency resolution of the spectrum analyser. The smaller the filter bandwidths, the greater the resolution.

Outputs of each filter are scanned by an electronic switching mechanism and used to define the controlling voltages across the vertical deflection plates of the CRT. A scan generator controls the electronic switching mechanism and provides the controlling voltages across the horizontal deflection plates. The displayed waveform is thus a collection of signal amplitudes, one for each bandpass filter, across the CRT screen.

A real-time spectrum analyser has the advantage that rapidly occurring events are detected and displayed – hence the name real-time. However, its main disadvantage is the large number of bandpass filters used. This restricts the use of a real-time spectrum analyser to measuring and displaying low frequency input signals, say, from DC to about 20 kHz.

Swept tuned spectrum analysers

The principle of using a bandpass filter to determine the amplitude of a signal component is also used in **swept tuned spectrum analysers.** However, only a single filter is used, the filter being tunable over a range of frequencies. *Figure 3* illustrates the concept. A scan generator is used, directly, to tune the bandpass filter over the frequency range under test. The amplitude of the signal output of the bandpass filter thus varies with time, according to its swept centre frequency.

Swept tuned spectrum analysers are much less expensive to produce than real-time spectrum analysers. However, it is difficult to make a tunable bandpass filter with sufficient accuracy and tuning range to realise a high performance spectrum analyser.

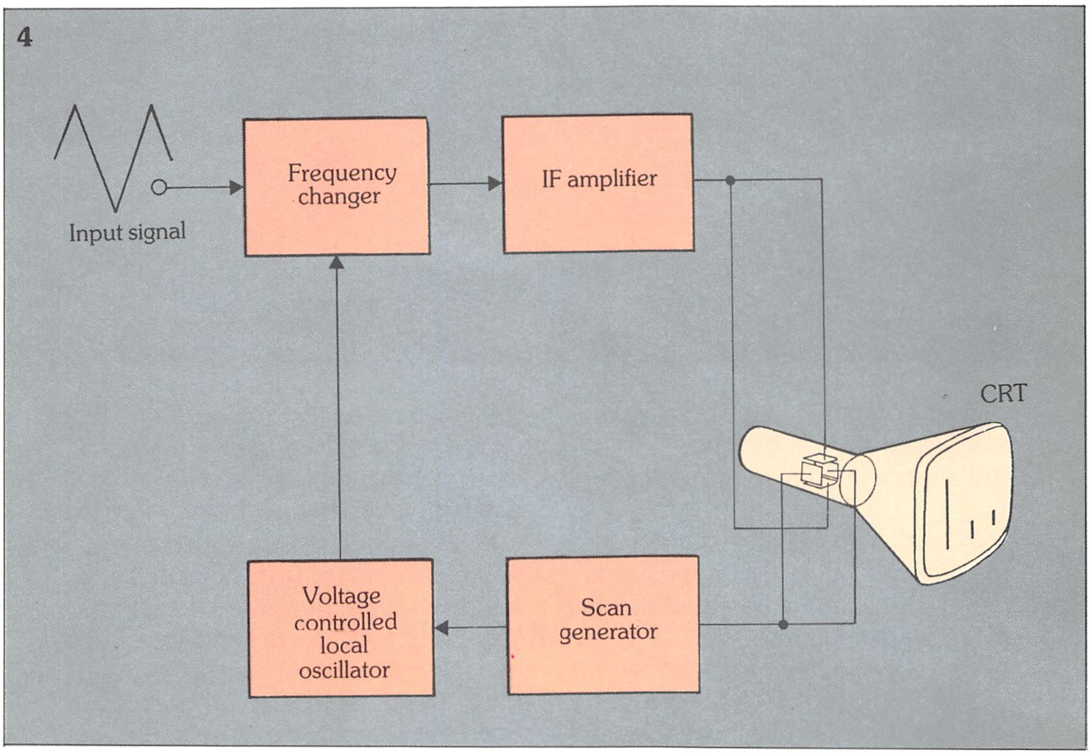
An extension of the swept tuned spectrum analyser principle is used in the **swept superheterodyne spectrum analyser**, where a tunable local oscillator beats with the incoming input signal to produce intermediate frequencies which are amplified by IF amplifiers. *Figure 4* shows the swept superheterodyne spectrum analyser principle.

A swept superhet spectrum analyser has many advantages: high sensitivity (thanks to the IF amplification); an extremely wide frequency range (many decades); and variable resolution. However, the display must be constructed from a number of parts of the total spectrum under observation, over a period of time. It does not, therefore, operate in real-time.

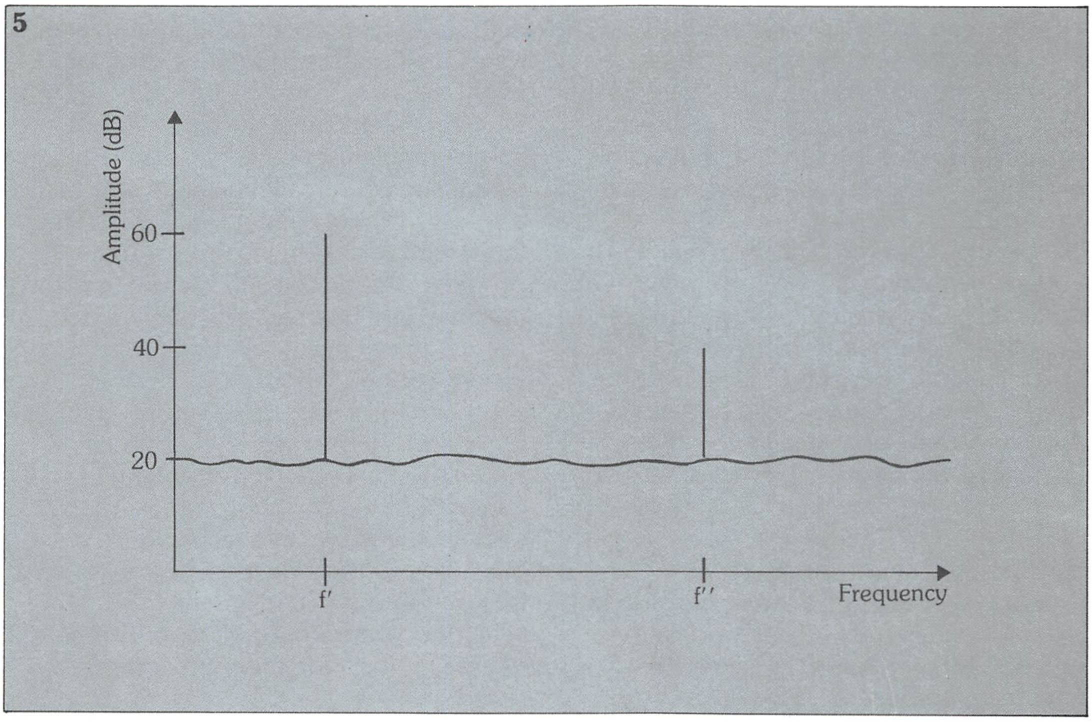
Fourier transform spectrum analyser

The third basic method used in spectrum analysers digitally processes the input signal under microprocessor control, to give frequency and amplitude data ready for display. The digital process is known as a **Fourier transformation** and so spectrum analysers operating on this principle are known as **Fourier transform spectrum analysers**. Frequency ranges of this type of spectrum analyser are, typically, from DC to about 100 kHz.

Digital processing of signals allows considerable control over the display. The user, for example, may store, erase, alter, move or compare with previous analyses.



4. Swept superhet spectrum analyser. A tunable local oscillator beats with the incoming input signal to produce intermediate frequencies, which then pass through IF amplification.



5. Display showing the use of a spectrum analyser to measure signal-to-noise ratio.

Spectrum analyser applications

We have already mentioned an example of the use of a spectrum analyser to detect and measure harmonically related distortion components in a signal. Non-harmonic related components, usually referred to as **spurious** components, are just as easily detected.

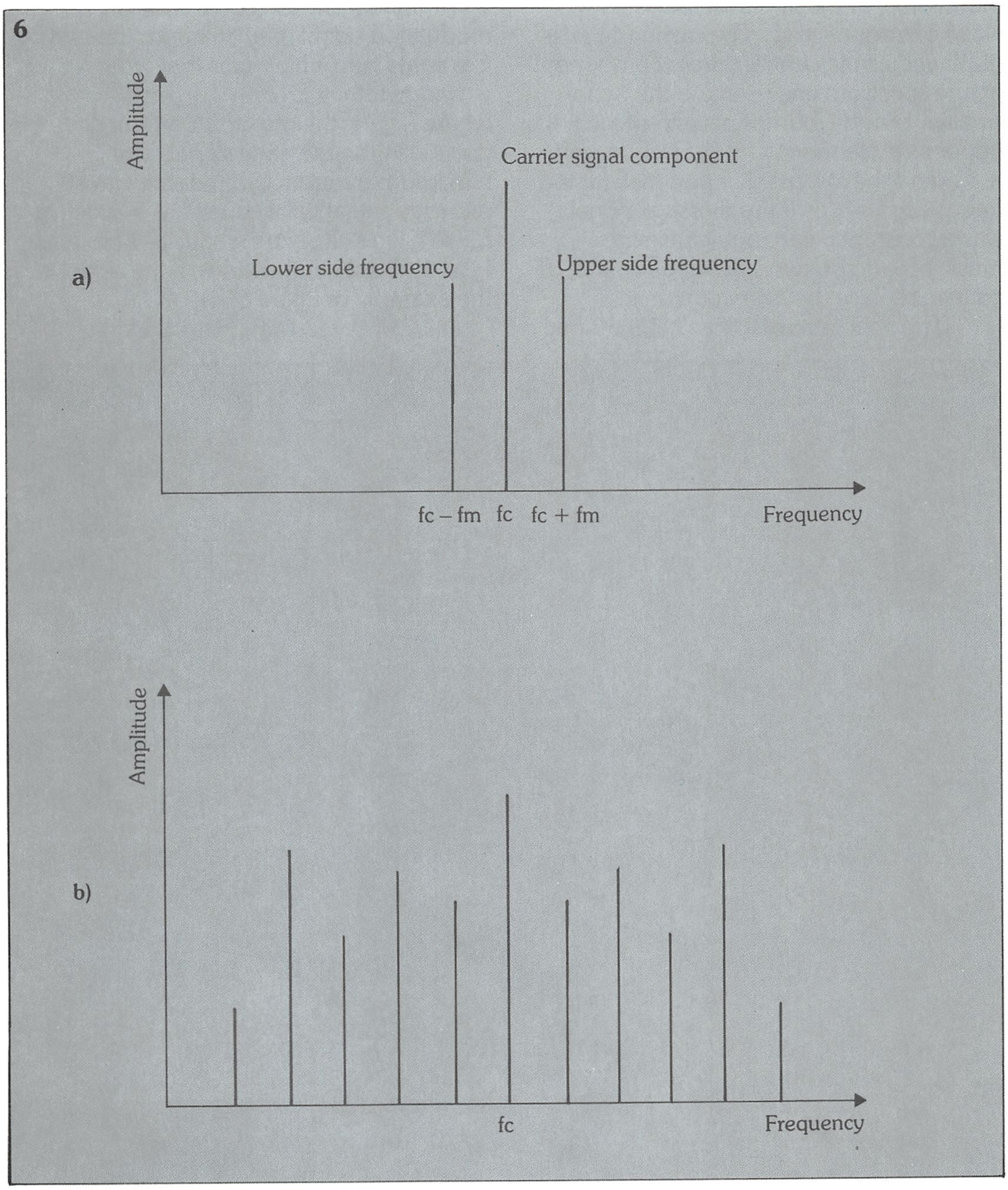
Another application of a spectrum analyser is to measure signal-to-noise ratio. *Figure 5* shows a possible display where the background noise is seen as a more or less constant band at the bottom of the screen. Two frequency components, at f′ and f″ are also shown. The vertical axis of the display is calibrated in decibels, so the signal-to-noise ratio may be simply obtained by subtracting the amplitude of the noise level from the amplitude of the chosen frequency component. The signal-to-noise ratio of the component f′, for example, is about:

$60 - 20 = 40$ dB

That of the component f″ is:

$40 - 20 = 20$ dB

6.(a) Spectrum of an envelope amplitude modulated signal; (b) spectrum of a frequency modulated signal.



Similar measurements of noise factor of a device may be taken with a spectrum analyser, by measuring the signal-to-noise ratios at the device's input and output. Since:

$$\text{noise factor} = \frac{\text{input signal-to-noise ratio}}{\text{output signal-to-noise ratio}}$$

the noise factor may then be calculated.

A spectrum analyser may be used to measure and display radio and television signals before transmission. *Figure 6a* shows the possible displayed spectrum of an envelope amplitude modulated signal. A carrier signal, of frequency f_c, is modulated with a single message sine wave of frequency f_m. The carrier signal is displayed as the central, larger component by the spectrum analyser and the two smaller, outside, components represent the upper **side frequency** (at the frequency $f_c + f_m$) and the lower side frequency (at the frequency $f_c - f_m$). If the message signal was a constantly varying signal, say, a music broadcast, two sidebands would, of course, replace the side frequencies.

The spectrum of a possible frequency modulated signal is shown in *figure 6b,* where a carrier, of frequency f_c, is modulated by a single message signal, of frequency f_m. Unlike the amplitude modulated signal's spectrum, the spectrum of *figure 6b* has a number of components around the carrier. The exact number depends on the deviation in frequency which the carrier undergoes due to modulation by the message signal. The deviation in frequency from the carrier is proportional to the amplitude of the modulating signal. Its maximum value is the maximum frequency deviation, Δf.

The bandwidth of a frequency modulated signal may be approximated by **Carson's rule** which says that:

$$\text{bandwidth} = 2\,(\Delta f + f_{max})$$

where f_{max} is the maximum frequency present in the message signal. The maximum frequency of the sine wave message signal is, of course, its frequency f_m, so if we define Δf as, say, 10 kHz, and f_m as 1 kHz the bandwidth of the signal in the example of *figure 6b* is:

$$2\,(10{,}000 + 2{,}000)\ \text{Hz} = 24\ \text{kHz}$$

Glossary

Carson's rule	a guide which defines the bandwidth of a frequency modulated signal as approximately $2\,(\Delta f + f_{max})$ where Δf is the maximum deviation of the carrier, and f_{max} is the maximum message frequency
Fourier transform	type of spectrum analyser in which the input signal is digitally processed as a Fourier transformation
frequency domain analysis	the type of analysis implied by the used of a spectrum analyser
real-time spectrum analyser	type of analyser in which the input signal is applied to a bank of bandpass filters to split up the signal into its individual frequency components
spectrum analyser	test equipment which allows measurement and display of a signal in the frequency domain, i.e. as a graph of amplitude against frequency
swept superhet spectrum analyser	type of spectrum analyser in which the incoming input signal beats with the output of a tunable local oscillator to produce intermediate frequencies
swept tuned spectrum analyser	type of spectrum analyser which uses a tunable bandpass filter to select frequency components present in the input signal
time domain	signal amplitude as a function of time

MICROPROCESSORS

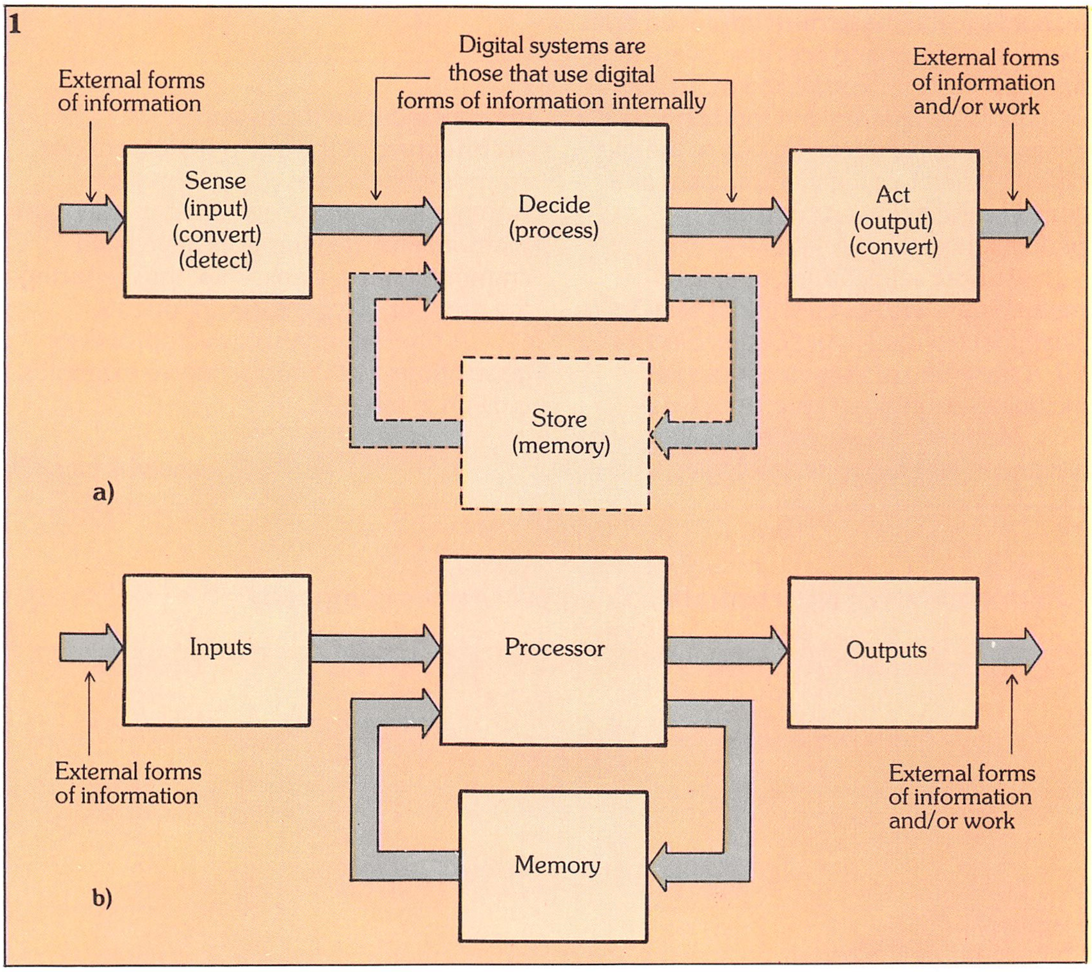
1 The universal chip

Introduction

Microprocessors are the units which perform both control and processing functions in computers and other digital systems, in applications as diverse as electronic toys and complex industrial control systems. Their versatility is due to the fact that they are programmable. Hundreds of thousands of microprocessors can be manufactured to the same specification, hence cutting manufacturing costs, but each can then be programmed to perform a different function.

In some applications, for example pocket calculators or digital watches, microprocessors are mass produced to a specific design. This reduces testing and assembly time, increases efficiency and hence reduces production costs. The very high development costs, though, are only justified in high quantity production runs.

It is the adaptability of the programmable microprocessor – the fact that systems can be modified, or changed completely, by programming – that makes it all pervasive in the modern world of electronics. In this respect it earns its name of the 'universal chip'.



1. The block diagram of the universal system organisation shown in **(a)**, is related to a microcomputer shown in **(b)**. The role of the microprocessor performing the 'decide' function is clear.

Universal system organisation

The by now familiar concept of the **universal system organisation** (see *Digital Electronics 1*) seen in *figure 1a* is applied to a microcomputer system in *figure 1b*. This illustrates that the microprocessor cannot function alone – it is one of the many building blocks in a computer system.

As you can see, the microprocessor performs the *decide* function; the input and output interfaces link the microprocessor to peripheral devices such as keyboards and printers; and the memory interface chip stores information. The microprocessor, therefore, is able to input information in the form of digital signals, process that information according to a stored program, and then output information in the form of digital signals.

Microprocessor technology

Since the introduction of the first microprocessor in 1971, it has been the advances made in IC manufacturing technology that have most improved both the performance and cost effectiveness of microprocessors. Today, the more than one hundred different microprocessors available are produced by one of the six different methods, each formulated as a compromise between price and performance.

As both semiconductor and IC manufacture are covered in detail in *Solid State Electronics 10* and *Digital Electronics 10*, a summary of these technologies including examples is given in *table 1*.

Along with these increasingly advanced fabrication technologies, improvements in the design or **architecture** of the microprocessor have considerably increased performance. Another factor in its 'evolution' is that more memory and I/O interfaces have been crammed onto a single chip, thus reducing the number of external components.

A brief summary of microprocessor generations, with representative examples, is given in *table 2*.

(continued in part 37)

Table 1
Microprocessor manufacturing technologies – a summary

			Examples
Bipolar	Unsaturated	ECL	Motorola 10800
		Schottky	Intel 3001 Advanced Micro Devices AM 2901
	Saturated	I^2L	TI SBP 0400 TI SBP 9900
MOS	CMOS		Intersil 6100 RCA COSMAC
	n-channel		Intel 8080, 8085, 8748, 8086 Motorola 6800, 6802, 6809 Fairchild F8 MOS Technology MCS 6502 Zilog Z80, Z8, Z8000
	p-channel		Intel 4004, 4040, 8008 National IMP PACE

Table 2
Performance of representative microprocessors compared

Microprocessor	Year introduced	Data path	Directly addressable memory (bytes)	Estimated relative throughput*
8008	1972	8-bit	16K	1
8080	1973	8-bit	64K	10
Z80	1976	8-bit	64K	20
8748	1977	8-bit	4K	20
8086	1978	16-bit	1024K	100
Z8000	1979	16-bit	8192K	100

* 8008 = 1